Optimizing rare variant association studies in theory and practice

Optimizing rare variant association studies in theory and practice

A dissertation presented

by

Ran Wang

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Genetics and Genomics

Harvard University

Cambridge, Massachusetts

April 2014

Dissertation Advisor: Dr. Joel N. Hirschhorn Ran Wang

Optimizing rare variant association studies in theory and practice

ABSTRACT

Genome-wide association studies (GWAS) have greatly improved our understanding of the genetic basis of complex traits. However, there are two major limitations with GWAS. First, most common variants identified by GWAS individually or in combination explain only a small proportion of heritability. This raises the possibility that additional forms of genetic variation, such as rare variants, could contribute to the missing heritability. The second limitation is that GWAS typically cannot identify which genes are being affected by the associated variants. Examination of rare variants, especially those in coding regions of the genome, can help address these issues. Moreover, several studies have recently identified low-frequency variants at both known and novel loci associated with complex traits, suggesting that functionally significant rare variants exist in the human population.

However, without sufficiently large sample size, we are underpowered to detect rare variant effects due to the low allele frequencies and the large numbers of rare variants in the exome. This dissertation is broadly divided into two parts to explore strategies for optimizing the power of rare variant association studies. First, we developed a cost-efficient pooled sequencing scheme as well as the analytic framework that ensures low false positive and false negative rates in variant discovery. We showed that this strategy is good for follow-up studies of candidate genes and for identifying potential genetic diagnosis in well-phenotyped patients. Second, we employed forward simulation to assess the usefulness of founder populations in rare variant

association studies and compare the efficiency of exome array genotyping *vs.* high coverage exome sequencing. We developed a novel simultaneous simulation of sequence variation in the non-Finnish European and the Finnish population that closely approximates the empirical sequence data. We showed that studies of founder populations like Finland can substantially increase power for discovery in a subset of genes and exome chip is currently much more cost-efficient than exome sequencing. Taken together, our results have highlighted the usefulness of having diverse sets of populations (ideally founder populations) and employing cost-efficient study designs such as exome chip followed by pooled sequencing to boost power of rare variant association studies.

TABLE OF CONTENTS

ACKNOWLEDGEMENT

First, I would like to thank my dissertation advisor, Joel Hirschhorn. When I first joined the lab, I was concerned about lack of experience in human genetics as well as insufficient computational skills. Thank you for encouraging me to pursue my own path. Over the four years, you have given me great freedom to explore my own research interest and have always been supportive. You showed me the utmost dedication, efficiency, and passion of a great scientist that I truly admire. The lessons I learnt from you are invaluable, and I will carry them with me throughout the rest of my life.

I would also like to thank all the past and present members of the Hirschhorn lab. Having spent so much time in the lab, you are like families to me. Thank you for creating a fun, supportive, and animated environment. Special thanks to Charleston Chiang, for mentoring me during my rotation and helping me get started with my projects, for stimulating discussions even after you left the lab, and for various lab traditions (such as game night, and all you can eat sushi dinner) you started. Thank you also to Rigel Chan and Elaine Lim, for being great classmates and going through graduate school together. I would also like to thank Andrew Dauber for all the great advice and for directing the first first-author paper of my life. Thank you, Thutrang Nguyen, Michael Turchin and Cameron Palmer, for making the lab so lively, which I enjoyed from the very beginning. Thank you, Sailaja Vedantam, for feeding me snacks all the time and for listening to my complaints. Thank you, Rany Salem, for bearing with the naughty graduate students and for printing out all the posters for us. Thank you, Tonu Esko and Tune Pers, for the European fashion you've brought to the lab. Finally, thank you, Jennifer Moon, for helping out with bench work and for organizing all the lab events.

I would also like to show my gratitude for the guidance of the members of my

dissertation advisory committee: Christine Seidman, Christopher Newton-Cheh, and Alkes Price. Thank you also to the members of my dissertation examining committee: Jose Florez, Daniel MacArthur, and Sohini Ramachandran, for accommodating my exam date among your busy schedules and for reading this dissertation.

A huge thank you to Min Lin, for being the great surprise of my life, and for making me feel loved all the time despite being over 9,000 miles apart and having 12 hour time difference. Thank you for always being there for me, no matter I am stressed about deadlines or feeling nervous about upcoming DAC meeting. I know I'm not the easiest person to love and to share a life with. Thank you for doing it anyway, and for wanting to do it anyway.

Finally, I would like to thank my parents, Chunxiang and Weiguo, for the unconditional support throughout the years. You are the best parents I could ever ask for! Thank you for bearing with me, especially when I am being difficult. Thank you for all your selfless sacrifices without which I could not be here today. Thank you, mom, for being a close friend and sharing my secrets. Thank you, dad, for loving me so much though you almost never say it. Even though you cannot be here for my defense and graduation ceremony, I hope that I have not let you down, and I dedicate my dissertation to you.

ATTRIBUTIONS

Chapter 2:

Sophie R. Wang: Together with my advisor and A. D., I participated in designing the experiments. I developed the analytic framework for dealing with pooled sequencing data and assessed data quality by comparing to whole exome sequencing data. I wrote relevant parts of the manuscript.

Heather Carmichael: Searched for HGMD mutations in the pooled sequencing data. Performed family analysis and looked through clinical phenotypes of carries of IGF1R mutations. Wrote relevant result sections of the manuscript.

Shayne F. Andrew: Performed IGF1R functional studies and wrote relevant parts of the manuscript.

Timothy C. Miller: Heavily involved in recruiting the short stature patients. Documented description of the cohort.

Jennifer E. Moon: Validated pooled sequencing variants via Sanger sequencing.

Michael A. Derr: Performed IGF1R functional studies and wrote relevant parts of the manuscript.

Vivian Hwa: Performed IGF1R functional studies and wrote relevant parts of the manuscript.

Joel N. Hirschhorn: Conceived of the overlapping pooling design. Provided helpful comments during all phases of the project.

Andrew Dauber: Conceived of, designed, and directed the project. Discussed and interpreted analysis results. Heavily edited the draft of the manuscript and wrote most of the discussion.

Chapter 3:

Sophie R. Wang: I participated in designing the sequencing experiment in the short stature patient cohort and FINRISK height extreme samples. Together with my advisor, I conceived of and designed the sequencing experiment in the height extreme samples from Estonian Biobank. Together with J.E.M., I designed and validated the primers for sequencing the Estonian height extreme samples. I conducted the analysis of all pooled sequencing data. I interpreted the results with my advisor and wrote the Chapter.

Heather Carmichael: Performed family analysis and looked through clinical phenotypes of NPR2 mutation carriers in the short stature patient cohort. Wrote relevant result sections.

Christina M. Jacobsen: Performed functional analyses of NPR2 mutations.

Timothy C. Miller: Heavily involved in recruiting the short stature patients.

Jennifer E. Moon: Validated pooled sequencing variants via Sanger sequencing. Designed and validated some of the primers for sequencing the Estonian height extreme samples.

Andrew Dauber: Conceived of, designed, and directed the sequencing experiment in the short stature patients cohort. Discussed and interpreted analysis results.

Joel N. Hirschhorn: Conceived of and directed the project. Interpreted the results and edited the chapter.

Chapter 4:

Sophie R. Wang: Together with my advisor, I conceived of and designed the project. I performed all analyses. I interpreted the results, wrote the manuscript, and edited the manuscript per reviewers' comments.

Vineeta Agarwala: Provided the model and parameters for simulating the European population. Provided helpful comments on using ForSim and on editing the manuscript.

Jason Flannick: Developed the model for simulating the European population along with V.A.. Provided helpful comments on the manuscript.

Charleston W.K. Chiang: Discussed and interpreted analysis results. Conceived of some analysis ideas. Provided helpful comments on the manuscript.

David Altshuler: Generated the phenotype residuals for the Nigerian cohort. Independently cleaned and conducted association analysis using the Nigerian cohorts. Cleaned the CNV dataset for the Nigerian cohort.

GoT2D Consortium: Provided the empirical whole exome sequencing data.

Joel N. Hirschhorn: Conceived of and designed the experiments. Critiqued and interpreted the analysis results and edited the manuscript.

Chapter 1

Introduction

A PREAMBLE

Genetic studies have revealed thousands of loci contributing to common polygenic human diseases and traits. Studies to date have mostly focused on studying individual common variants, because they could be more readily assayed with initial genomic technologies. Nonetheless, the genetic variants discovered thus far typically explain only a small fraction of the estimated heritability. Some of the so-called missing heritability is likely due to additional common variants yet to be discovered. On the one hand, studies have been limited by sample size; those with larger sample sizes continue to reveal many new loci[1–3]. On the other hand, indirect statistical methods indicate that common variants collectively capture at least 30% (and likely more) a large proportion of the heritability for a number of diseases and traits[4–6].

The additional sources of missing heritability remain unclear. One hypothesis is that much of the missing heritability is due to rare genetic variants. The theoretical case for an important role of rare variants is that alleles predisposing to disease are likely to be deleterious and thus kept at low frequencies by purifying selection[7–9]. Because rare variants are too numerous and occur too infrequently, rare variant association studies (RVAS), like common variant association studies (CVAS)[10], would require large sample collections and careful statistical analysis to achieve adequate power to detect genes underlying diseases.

In reality, the practice of RVAS is still in its infancy. Some early efforts were premised on the notion that rare variants underlying common diseases could be readily identified in small numbers of samples. The few discoveries from RVAS thus far are mostly from candidate gene studies rather than unbiased gene discovery and, in some cases, reach only nominal rather than genome-wide levels of significance[11–18]. The analytical methodology for RVAS remains in flux, although many groups have proposed a rich collection of possible statistics[19–21].

To address these challenges, in this dissertation we present an initial attempt to evaluate the design of RVAS and explore strategies for maximizing power of RVAS. In Chapter 2, we will describe a cost-efficient pooled sequencing scheme, which is further illustrated by applications of it in both Chapter 2 and Chapter 3. In Chapter 4, we present an analytical framework for evaluating the design of RVAS in different study populations. As the field of human genetics moves forward to explore expanded sources of variation in more diverse populations, we believe our approach will be useful to guide future studies.

However, before describing the results and implications of each chapter, a more thorough introduction to human genetic analysis is warranted. As such, the remainder of this chapter is broadly divided into three sections. The first section provides an overview of complex trait genetics, including the approaches for mapping genes underlying a phenotype of interest. The second section reviews the approaches for studying human complex traits, describing the process from early linkage studies to genetic association studies of both common and rare variants. The third section is a brief introduction to pooled sequencing and its limitations and challenges.

GENETICS OF COMPLEX TRAITS

Interest in the genetic basis of disease originated with the rediscovery of Mendel's laws in the beginning of the 20th century. Subsequent studies have identified many of the genes responsible for Mendelian diseases, conditions that are strongly influenced by highly penetrant variants in a single locus and follow a clear familial pattern. While Mendelian traits formed the basis for classic genetics, it has become clear that most common human traits and diseases are complex traits, influenced by many genetic loci, along with environmental factors. Examples include cardiovascular, metabolic and neuropsychiatric disorders.

While most alleles underlying Mendelian diseases are rare and relatively new in the population, controversy exists regarding the likely allelic spectrum of variants contributing to complex traits. Two major lines of reasoning, representing two extremes of the frequency spectrum of variants, exist: the "common disease common variant" hypothesis and the "common disease rare variant" hypothesis.

The common disease common variant hypothesis

At one extreme, the "common disease common variant" hypothesis (CDCV hypothesis) posits that common genetic variants underlie susceptibility to most common traits. This hypothesis is rooted in the assumption that the majority of polymorphisms predisposing to complex diseases are not evolutionarily deleterious and thus can rise to high frequencies. Several evolutionary scenarios might explain why disease-causing variants escape purifying selection, such as positive or balancing selection[22], a late disease onset, and changing direction of selection[23]. To date, genome-wide association studies have successfully identified thousands of common variants[24], which to a certain extent supports the CDCV hypothesis. Examples of such common variants include the APOE ε4 allele in Alzheimer's disease[25] and PPARγ Pro12Ala in type II diabetes[26].

The common disease rare variant hypothesis

On the other extreme is the "common disease rare variant" hypothesis (CDRV hypothesis). This model dictates that phenotypic variation in complex traits is caused by moderately highly penetrant rare variants. The key assumption here is that the majority of disease-causing mutations are also mildly evolutionarily deleterious and thus kept at low frequencies by

purifying selection. A high mutation rate counterbalances the action of purifying selection and determines the cumulative frequency of disease-causing variants[9]. A well-known example supporting this hypothesis is the BRCA1 and BRCA2 breast cancer susceptibility mutations. Women carrying mutations in these genes have a lifetime risk of breast cancer as high as 80%. Thousands of mutations have been found within these two loci; they are collectively quite common but individually rare or even private to the family[27,28].

Both models have been supported by earlier theoretical studies[9,29,30]. The difference in conclusions is mostly due to difference in the rate of deleterious mutations and the strength of selection against new mutations[7]. In reality, the dichotomy between the two models might not be absolute. It is likely that both common and rare variants, along with interactions between variants both common and rare and interactions of genetic variants with the environment, contribute to complex traits. Neither model is likely to be exclusively accurate for any trait in general, and the genetic architecture and allelic spectrum of causal variants is likely to differ from trait to trait.

APPROACHES FOR STUDYING HUMAN COMPLEX TRAITS

Having introduced the hypotheses regarding the genetic basis of complex traits, we next discuss the approaches that have been used to map genetic loci responsible for phenotypic variation. These approaches can be broadly grouped into two categories: linkage analysis and genetic association studies. Linkage analysis, a statistical technique that successfully identified casual genes for many Mendelian disorders, has had limited success for gene mapping of complex traits. An alternative is genetic association studies, which are analogous to traditional epidemiologic association studies. Instead of seeking association between traditional risk

variables and disease outcome, a genetic association study looks for an association between a genetic variant and a specified condition.

Limited success of linkage analysis for complex traits

The principle of linkage analysis is founded on the co-segregation of chromosomal regions identical by descent along with the disease phenotype of interest within large families. Genome-wide linkage studies became feasible in the 1980s, with genome-wide linkage map of hundreds of DNA sequence variations[31]. Since the first successful demonstration of systematic linkage analysis to localize the gene causing Huntington disease in 1983[32], linkage analysis has led to the discovery of causal mutations for a large number of Mendelian disorders[33].

Genome-wide linkage analysis has also attempted on a large number of complex traits. Despite the discovery of some clearly relevant loci that contribute to susceptibility of complex diseases such as inflammatory bowel disease[34,35] and type 1 diabetes[36], most of these studies have failed to identify a locus by strict criteria[37] and the results of studies of the same disease are often inconsistent[38]. The reasons for this lack of success can be ascribed to a few basic problems. First, mutations in any one of multiple genes may result in identical phenotypes (locus heterogeneity). Genetic heterogeneity hampers linkage analysis, because a chromosomal region may co-segregate with a disease in some families but not in the others. Second, some traits may require multiple variants to act in concert (polygenic inheritance). Polygenic inheritance would complicate linkage analysis because no single locus is strictly required to cause a disease. Third, loci underlying complex traits each likely have much smaller effects (*i.e.*, reduced penetrance), each potentially also interacting with other genes or environment. Thus, the genotype at a given locus may affect the probability of disease, but not fully determine the outcome. In such cases,

the signal-to-noise ratio is reduced in a linkage analysis, making it harder to pinpoint a linked region[38–42]. Moreover, linkage analysis has very low resolution. Even if a chromosomal region can be definitively mapped, the linkage peak region often covers several megabases. Therefore, extensive candidate gene studies are still required to narrow analysis to the causal gene or genes in the linked region[42,43].

The principles of association studies

An alternative approach to identify the genetic basis of complex traits is population-based association studies. Association studies assess the correlation between genetic markers and trait differences among unrelated population samples. The aim is to track the small chunk of ancestral chromosome where the causal mutation first arose but have decayed over time due to recombination. A higher frequency of a genetic marker in individuals with a disease can be interpreted as meaning that the tested marker increases disease risk. Genetic markers and trait can also become associated by other mechanisms, among which linkage disequilibrium (LD) is of greatest interest. The human genome is composed of long "block-like" regions where SNPs are non-randomly associated with one another, a phenomenon known as LD[44–46]. If the genotyped marker is in LD with the causal variant, the genotyped marker would also appear to be associated with the phenotype, and allows localization of the genetic locus. In essence, association studies and linkage analysis are quite similar, both relying on the co-segregation of chromosomal regions that contain both the genetic marker and the disease locus. They differ in that linkage analysis tracks chromosomal regions over a few generations of recent ancestry, and association studies track smaller chromosomal regions over many generations of historical ancestry.

Despite the similarity in fundamental principles, association studies have greater statistical power to detect relatively common alleles with modest effect sizes than linkage analysis, requiring fewer (although still a large number) individuals in the study[47,48]. The large number of individuals required for a powerful study design in either scenario also means that it is easier to conduct a powerful association study as it is relatively easier to recruit unrelated affected individuals than to collect large numbers of pedigrees each with multiple affected individuals, particularly for diseases of old age[43]. In addition, association studies have higher resolution than linkage analysis, because the region around a marker shared identically by descent in unrelated individuals will be much smaller than the shared region between family members[43,47].

There are different ways in which association studies can be categorized. Regarding the trait of interest, there are case-control studies for dichotomous phenotypes and population-based studies for continuous quantitative traits. Regarding the scale of the study, there are two broad categories: candidate-gene studies and genome-wide studies. Regarding the frequency spectrum of tested variants, association studies could evaluate common variants or rare variants. Association studies of individual common variants are named as genome-wide association studies (GWAS), whereas association studies of rare variants are often described as resequencing studies. This nomenclature conflates statistical methodology (association testing) and laboratory methodology (DNA sequencing). We will use the terms from Zuk *et al*.[10]: common variant association study (CVAS) and rare variant association study (RVAS).

CVAS: from candidate gene approach to genome-wide studies

Association studies to date have largely focused on studying individual common variants. Early CVAS had typically taken a candidate gene approach. Candidate genes are often selected

based on potential biological relevance or from prior human genetic evidence[42,43]. Such studies have successfully identified a number of genes that contribute to susceptibility to common disease, such as *PPARG* and *KCNJ11* for type 2 diabetes[26,49] and *ABCA1*, *APOA1*, and *LCAT* for plasma levels of high-density lipoprotein cholesterol[11]. Nonetheless, the inherent problem with candidate gene approach is that it is limited by our understanding of the disease pathophysiology. Moreover, the significant findings of association in candidate-gene studies, which were typically not held to the current genome-wide threshold of significance ($p<5x10^{-8}$) are often not consistently associated with disease across a large number of independent studies[50,51].

Aided by the dense genetic map from the HapMap project[46] and the development of highly accurate, cost-efficient, SNP array technology, the paradigm in CVAS has largely shifted from candidate gene studies to genome-wide studies over the last few years. The basic scheme used is to catalogue very common variants and genotype them either directly or indirectly (through LD). Because no assumptions are made about the genomic location of the causal variants, the genome-wide approach could exploit the strengths of association studies without having to guess the identity of the causal genes. Today, millions of SNPs can be assayed by commercialized SNP arrays simultaneously. By properties of LD, these SNPs quite adequately cover most of the common variation in human populations, particularly for European-derived and East Asian populations due to their extended LD blocks[52,53]. Moreover, the remaining catalogued common variants not covered by the commercial arrays can be recovered by computational approaches known as imputation.

The success and failure of CVAS

To date, CVAS have successfully identified thousands of common variants associated with

hundreds of diseases and traits[54]. These studies have "rediscovered" many genes that have been shown to be important. For example, of the 95 loci found to be associated with blood lipid levels, 18 genes were previously implicated in Mendelian lipid disorders, and several others had been known to influence lipid metabolism[55]. These studies have also highlighted biological pathways previously not known to be relevant to a particular disease or trait. The loci associated with Crohn's disease point unambiguously to autophagy and interleukin-23-related pathways[56], and the height loci include genes encoding chromatin proteins and hedgehog signaling[57]. Finally, many new identified loci do not implicate genes with known functions. For these findings, greater effort will be required to generate hypotheses for future work, but such efforts could open new avenues of biological research of complex traits. For instance, a ubiquitin ligase, *MYLIP*, had no recognizable role in lipid metabolism before CVAS, but has since been shown to regulate cellular LDL receptor levels[3,58].

Despite the widespread success of CVAS in identifying genes and pathways relevant to complex traits, there are two major limitations. First, common variants identified by CVAS almost universally have small effects (~1.1 to 1.5 fold increased risk), and in combination explain a small proportion of the phenotypic variance attributable to genetic causes (the "heritability")[59–61]. In the case of height, where heritability is as high as 80%, only ~10% of phenotypic variance is explained by currently published CVAS association signals[2]. Though using the mixed linear modeling approach over 50% of the phenotype can be attributed to common variants[5], it can be argued that common variants do not capture the full range of genetic variance and that part of the missing heritability will need to be addressed with rare variants and broad sense heritability components such as gene-gene or gene-environment interactions[62–64]. Second, CVAS likely have not identified the causal SNPs, but rather have identified variants that

are proxies for the causal SNP or haplotype. Association signals often do not point unambiguously to a particular gene.

The biologic pictures being revealed by CVAS are still quite incomplete. We should strive for as complete a catalogue of validated risk variants as possible. On the one hand, additional CVAS, in larger samples and multiple ethnicities, will continue to reveal many more new loci[1–3] and follow-up by approaches including fine mapping, genomic analysis of gene expression in human tissues, and screen for mutations on marker-containing haplotypes will lead to better understanding of the results. On the other hand, complementary approaches such as RVAS will look into additional sources of missing heritability.

RVAS: sequencing and complementary approaches

The frequency boundaries for defining rare variants in the literature vary. Here, we use minor allele frequency of less than 5%, as variants of this frequency range are poorly captured by the commercial arrays for genome-wide CVAS. Genome-wide surveys for such variants will eventually be carried out in a manner similar to CVAS, with very large sample sizes that will provide sufficient statistical evidence to implicate variants on the basis of association evidence alone.

With the arrival of the next-generation sequencing technologies, it has come into reach to have full genomic sequences available for multiple individuals rather than relying on a more or less representative fraction of the genome. High-coverage whole-genome sequencing will be the method of choice once it becomes more affordable, with the rapid increase in the sequencing capacity of existing platforms, as well as the development of new, less-expensive platforms. However, in the interim, it will be important to focus on more cost-effective alternative strategies.

One popular approach is whole-exome sequencing. Because the most obvious disease-influencing variants will be the clearly functional ones, whole-exome sequencing is a much smaller, more cost-effective and more easily interpreted bundle. An even cheaper alternative approach is to use genotyping arrays which account for less common variants, an example of which is exome chip. In Chapter 4, we present a simulation-based work where we compared the power and cost-efficiency of exome chip against exome sequencing. Imputation of rare variants into samples with existing genotype data is another likely complementary approach for future studies, especially as reference panels for imputation become larger and represent more populations.

Strategies for optimizing the power of RVAS

Until sequencing studies are inexpensive enough to for large sample sizes, it will be important to focus on designs that are optimized to detect the role of casual variants in smaller samples. Such designs may accelerate progress, by enabling early discoveries. Applications of these designs can be seen in Chapter 2 and 3.

First, isolated populations resulting from recent bottlenecks should make it easier to detect a subset of genes. Examples include Finland, Iceland, Ashkenazi Jews, Amish, Bedouins, and various endogamous groups in India. Studies across multiple such populations could prove valuable. In Chapter 4, we explore in depth the implication of founder effect in RVAS.

Second, extreme-trait sequencing is likely to be another important engine of discovery. A basic extreme-trait design would be to sequence a small, carefully selected population at one or both ends of the extremes of a phenotype. Because variants that contribute to the trait will be enriched in frequency in the extreme individuals, even small sample sizes may suggest candidate

variants that can then be genotyped for confirmation in a much larger collection of samples. For the follow-up of variants identified in extreme-trait sequencing, family members of the extreme individuals will be invaluable for confirming potentially causal variants through co-segregation analysis.

Third, early signals may be provided by the study of gene sets that are likely to be enriched for disease-associated loci. The best sets may consist of genes implicated by CVAS. Whereas CVAS and RVAS are sometimes thought of as alternatives, they are likely to be complementary. Targeted sequencing of pooled samples is well suited for studying candidate gene sets, which we will discuss in more detail below.

POOLED SEQUENCING

Motivations for pooled sequencing

As opposed to the naïve approach that aims at sequencing large regions in a single individual, there are many biological applications wherein the sequence of a small region must be determined from many independent samples. Examples might include studies of variation within specific regions from a large population, looking for associations between variants and trait in areas that might have been narrowed by prior linkage analysis, CVAS, or even medical diagnostics. Use of next-generation sequencing technologies for interrogating sequence for any of these tasks is hampered if libraries must be constructed independently from each sample or if different barcodes must be applied to each sample prior to sequencing, particularly if sample numbers run into many thousands.

The actual cost of sequencing a sample consists of two parts. The first part is the cost of preparing a DNA sample for sequencing which is referred to as library preparation cost. Library

preparation is also the most time-consuming and labor-intensive part of a sequencing study. The second part is the cost of the actual sequencing, which is proportional to the amount of sequence. The dramatic increase in the efficiency of the sequencing technology makes the costs of the sequencing step negligible for small target regions. Thus the main remaining cost is the sample preparation step.

This limitation raises the need for the development of multiplexing strategies that allow the processing of multiple samples per single sample preparation step at the cost of requiring additional sequencing capacity. Pooled genotyping has been used to quantify previously identified variations and study allele frequency distributions[65–67] in populations. Given an observed number of alleles and an estimate of the number of times an allelic region was sampled in the pool, it is possible to infer the frequency of the allele in the pooled individuals being studied. Pooled sequencing can be used to reach similar ends, as a strategy to cost-effectively capture all variation in a target region. Such an approach allows efficient use of next-generation sequencing technologies, as sequencing a large pool of individuals simultaneously keeps the number of redundant DNA reads low.

Overview of pooling methods

The basic idea behind pooled sequencing is that DNA from multiple individuals is pooled together into a single DNA mixture which is then prepared as a single library and sequenced. In this approach, the library preparation cost is reduced because one library is prepared per pool instead of one library per sample.

Pooling methods can be split into two categories. The first category puts each individual in only one pool. The naïve, disjoint pooling scheme offers insight into allele frequencies, but does

not offer the identity of an allele carrier. These types of methods are referred to as non-overlapping pool methods. The second category puts each individual in multiple pools and uses this information to recover each individual's genotype. These methods are referred to as overlapping pool methods. Many groups have developed sophisticated overlapping pooling designs, which encode the identity of each sample within the pooling pattern[68–70]. However, these designs usually involve complicated DNA pooling step. In our work, we developed a simple overlapping pooling design, which only allows decoding of singleton variants. But as we will show in Chapter 2 and 3, the singleton variants capture well the minor allele frequency range we aim to target in our studies.

Limitations and challenges of pooled sequencing

One obvious limitation of pooled sequencing is that haplotype information is not available. But this will be outweighed by the increased efficiency in studies where LD information is not as critical. Other major limitations include the ambiguous identification of individual carriers, especially in the presence of errors and difficulty in discerning whether a singleton variant is homozygous or heterozygous in an individual subject.

SNP calling and estimating the frequency of the minor allele from pooled samples, is a subtle exercise for at least three reasons. First, sequencing errors may have a much larger relevance than in individual SNP calling. While their impact in individual sequencing can be reduced by setting a minimum number of reads per allele, this would have a strong and undesired effect in pools because it is unlikely that alleles at low frequency in the pool will be read many times[71]. Second, an obvious source of error in the pooling approach is the unequal representation of each sample's DNA in the pool. This unequal representation could be due to human or

machine error. In experiments that rely on PCR amplification, the heterogeneity can be expected to be particularly strong. Individuals for which a larger DNA amount has been included in the pool will be overrepresented, which potentially causes a change in allele frequency estimates, while unrepresented samples in pool might lead to false negative discoveries[72].

In Chapter 2, we present a simple matrix pooling design that overcomes some of these obstacles. Under this design, each sample was sequenced in 2 pools (1 row pool and 1 column pool), which allows us to identify individuals carrying singleton variants. Moreover, the matrix design is less affected by unequal representation of each sample's DNA in the pool and enables us to filter out a lot of false positives from sequencing errors or other sources. In Chapter 3, we applied this approach to perform an RVAS of a particular gene (NPR2) where heterozygous mutations had been proposed to contribute to short stature.

SUMMARY

This introductory chapter hopefully has described and contrasted the major thoughts in human genetics research over the last 10-15 years, with a focus on using association studies to map disease genes. Aiming to provide the relevant background for the rest of this dissertation, we have compared and contrasted the common disease common variant and common disease rare variant hypotheses and discussed the methodology used to conduct an association study. We introduced the methodology of pooled sequencing and discussed its relative merits, setting the stage for the studies in Chapter 2 and 3. We also covered various strategies and designs for optimizing RVAS, which were evaluated under a simulation framework in Chapter 4.

REFERENCES

1. Van der Harst, P., Zhang, W., Mateo Leach, I., Rendon, A., Verweij, N., Sehmi, J., Paul, D.S., Elling, U., Allayee, H., Li, X., et al. (2012). Seventy-five genetic loci influencing the human red blood cell. Nature *492*, 369–375.

2. Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature *467*, 832–838.

3. Teslovich, T.M., Musunuru, K., Smith, A. V, Edmondson, A.C., Stylianou, I.M., Koseki, M., Pirruccello, J.P., Ripatti, S., Chasman, D.I., Willer, C.J., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. Nature *466*, 707–713.

4. Lee, S.H., DeCandia, T.R., Ripke, S., Yang, J., Sullivan, P.F., Goddard, M.E., Keller, M.C., Visscher, P.M., and Wray, N.R. (2012). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. Nat. Genet. *44*, 247–250.

5. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.

6. Yang, J., Manolio, T.A., Pasquale, L.R., Boerwinkle, E., Caporaso, N., Cunningham, J.M., de Andrade, M., Feenstra, B., Feingold, E., Hayes, M.G., et al. (2011). Genome partitioning of genetic variation for complex traits using common SNPs. Nat. Genet. *43*, 519–525.

7. Kryukov, G. V, Pennacchio, L.A., and Sunyaev, S.R. (2007). Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am. J. Hum. Genet. *80*, 727–739.

8. Goldstein, D.B., Allen, A., Keebler, J., Margulies, E.H., Petrou, S., Petrovski, S., and Sunyaev, S. (2013). Sequencing studies in human genetics: design and interpretation. Nat. Rev. Genet. *14*, 460–470.

9. Pritchard, J.K. (2001). Are rare variants responsible for susceptibility to complex diseases? Am. J. Hum. Genet. *69*, 124–137.

10. Zuk, O., Schaffner, S.F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M.J., Neale, B.M., Sunyaev, S.R., and Lander, E.S. (2014). Searching for missing heritability: Designing rare variant association studies. Proc. Natl. Acad. Sci. *111*, E455–64.

11. Cohen, J.C., Kiss, R.S., Pertsemlidis, A., Marcel, Y.L., McPherson, R., and Hobbs, H.H. (2004). Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science *305*, 869–872.

12. Johansen, C.T., Wang, J., Lanktree, M.B., Cao, H., McIntyre, A.D., Ban, M.R., Martins, R.A., Kennedy, B.A., Hassell, R.G., Visser, M.E., et al. (2010). Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. Nat. Genet. *42*, 684–687.

13. Ahituv, N., Kavaslar, N., Schackwitz, W., Ustaszewska, A., Martin, J., Hebert, S., Doelle, H., Ersoy, B., Kryukov, G., Schmidt, S., et al. (2007). Medical sequencing at the extremes of human body mass. Am. J. Hum. Genet. *80*, 779–791.

14. Rivas, M.A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C.K., Boucher, G., Ripke, S., Ellinghaus, D., Burtt, N., et al. (2011). Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat. Genet. *43*, 1066–1073.

15. Romeo, S., Yin, W., Kozlitina, J., Pennacchio, L.A., Boerwinkle, E., Hobbs, H.H., and Cohen, J.C. (2009). Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. J. Clin. Invest. *119*, 70–79.

16. Ji, W., Foo, J.N., O'Roak, B.J., Zhao, H., Larson, M.G., Simon, D.B., Newton-Cheh, C., State, M.W., Levy, D., and Lifton, R.P. (2008). Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nat. Genet. *40*, 592–599.

17. Diogo, D., Kurreeman, F., Stahl, E.A., Liao, K.P., Gupta, N., Greenberg, J.D., Rivas, M.A., Hickey, B., Flannick, J., Thomson, B., et al. (2013). Rare, low-frequency, and common variants in the protein-coding sequence of biological candidate genes from GWASs contribute to risk of rheumatoid arthritis. Am. J. Hum. Genet. *92*, 15–27.

18. Bonnefond, A., Clément, N., Fawcett, K., Yengo, L., Vaillant, E., Guillaume, J.-L., Dechaume, A., Payne, F., Roussel, R., Czernichow, S., et al. (2012). Rare MTNR1B variants impairing melatonin receptor 1B function contribute to type 2 diabetes. Nat. Genet. *44*, 297–301.

19. Basu, S., and Pan, W. (2011). Comparison of statistical tests for disease association with rare variants. Genet. Epidemiol. *35*, 606–619.

20. Stitziel, N.O., Kiezun, A., and Sunyaev, S. (2011). Computational and statistical approaches to analyzing variants identified by exome sequencing. Genome Biol. *12*, 227.

21. Ladouceur, M., Dastani, Z., Aulchenko, Y.S., Greenwood, C.M.T., and Richards, J.B. (2012). The empirical power of rare variant association methods: results from sanger sequencing in 1,998 individuals. PLoS Genet. *8*, e1002496.

22. Charlesworth, D. (2006). Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet. *2*, e64.

23. Di Rienzo, A., and Hudson, R.R. (2005). An evolutionary framework for common diseases: the ancestral-susceptibility model. Trends Genet. *21*, 596–601.

24. Hindorff, L., MacArthur, J., Morales, J., Junkins, H., Hall, P., Klemm, A., and Manolio, T. A Catalog of Published Genome-Wide Association Studies.

25. Corder, E.H., Saunders, A.M., Strittmatter, W.J., Schmechel, D.E., Gaskell, P.C., Small, G.W., Roses, A.D., Haines, J.L., and Pericak-Vance, M.A. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science *261*, 921–923.

26. Altshuler, D., Hirschhorn, J.N., Klannemark, M., Lindgren, C.M., Vohl, M.C., Nemesh, J., Lane, C.R., Schaffner, S.F., Bolk, S., Brewer, C., et al. (2000). The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. Nat. Genet. *26*, 76–80.

27. Walsh, T., Lee, M.K., Casadei, S., Thornton, A.M., Stray, S.M., Pennil, C., Nord, A.S., Mandell, J.B., Swisher, E.M., and King, M.-C. (2010). Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. Proc. Natl. Acad. Sci. U. S. A. *107*, 12629–12633.

28. McClellan, J., and King, M.-C. (2010). Genetic heterogeneity in human disease. Cell *141*, 210–217.

29. Pritchard, J.K., and Cox, N.J. (2002). The allelic architecture of human disease genes: common disease-common variant...or not? Hum. Mol. Genet. *11*, 2417–2423.

30. Reich, D.E., and Lander, E.S. (2001). On the allelic spectrum of human disease. Trends Genet. *17*, 502–510.

31. Donis-Keller, H., Green, P., Helms, C., Cartinhour, S., Weiffenbach, B., Stephens, K., Keith, T.P., Bowden, D.W., Smith, D.R., and Lander, E.S. (1987). A genetic linkage map of the human genome. Cell *51*, 319–337.

32. Gusella, J.F., Wexler, N.S., Conneally, P.M., Naylor, S.L., Anderson, M.A., Tanzi, R.E., Watkins, P.C., Ottina, K., Wallace, M.R., and Sakaguchi, A.Y. A polymorphic DNA marker genetically linked to Huntington's disease. Nature *306*, 234–238.

33. Jimenez-Sanchez, G., Childs, B., and Valle, D. (2001). Human disease genes. Nature *409*, 853–855.

34. Hugot, J.P., Chamaillard, M., Zouali, H., Lesage, S., Cézard, J.P., Belaiche, J., Almer, S., Tysk, C., O'Morain, C.A., Gassull, M., et al. (2001). Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature *411*, 599–603.

35. Ogura, Y., Bonen, D.K., Inohara, N., Nicolae, D.L., Chen, F.F., Ramos, R., Britton, H., Moran, T., Karaliuskas, R., Duerr, R.H., et al. (2001). A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature *411*, 603–606.

36. Nisticò, L., Buzzetti, R., Pritchard, L.E., Van der Auwera, B., Giovannini, C., Bosi, E., Larrad, M.T., Rios, M.S., Chow, C.C., Cockram, C.S., et al. (1996). The CTLA-4 gene region of chromosome 2q33 is linked to, and associated with, type 1 diabetes. Belgian Diabetes Registry. Hum. Mol. Genet. *5*, 1075–1080.

37. Lander, E., and Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat. Genet. *11*, 241–247.

38. Altmüller, J., Palmer, L.J., Fischer, G., Scherb, H., and Wjst, M. (2001). Genomewide scans of complex human diseases: true linkage is hard to find. Am. J. Hum. Genet. *69*, 936–950.

39. Lander, E.S., and Schork, N.J. (1994). Genetic dissection of complex traits. Science *265*, 2037–2048.

40. Chakravarti, A. (1999). Population genetics--making sense out of sequence. Nat. Genet. *21*, 56–60.

41. Risch, N.J. (2000). Searching for genetic determinants in the new millennium. Nature *405*, 847–856.

42. Hirschhorn, J.N., and Daly, M.J. (2005). Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. *6*, 95–108.

43. Carlson, C.S., Eberle, M.A., Kruglyak, L., and Nickerson, D.A. (2004). Mapping complex disease loci in whole-genome association studies. Nature *429*, 446–452.

44. Slatkin, M. (2008). Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. Nat. Rev. Genet. *9*, 477–485.

45. Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., et al. (2002). The structure of haplotype blocks in the human genome. Science *296*, 2225–2229.

46. International HapMap Consortium (2005). A haplotype map of the human genome. Nature *437*, 1299–1320.

47. Cardon, L.R., and Bell, J.I. (2001). Association study designs for complex diseases. Nat. Rev. Genet. *2*, 91–99.

48. Risch, N., and Merikangas, K. (1996). The future of genetic studies of complex human diseases. Science *273*, 1516–1517.

49. Florez, J.C., Burtt, N., de Bakker, P.I.W., Almgren, P., Tuomi, T., Holmkvist, J., Gaudet, D., Hudson, T.J., Schaffner, S.F., Daly, M.J., et al. (2004). Haplotype structure and genotype-phenotype correlations of the sulfonylurea receptor and the islet ATP-sensitive potassium channel gene region. Diabetes *53*, 1360–1368.

50. Ioannidis, J.P., Ntzani, E.E., Trikalinos, T.A., and Contopoulos-Ioannidis, D.G. (2001). Replication validity of genetic association studies. Nat. Genet. *29*, 306–309.

51. Hirschhorn, J.N., Lohmueller, K., Byrne, E., and Hirschhorn, K. A comprehensive review of genetic association studies. Genet. Med. *4*, 45–61.

52. Pe'er, I., de Bakker, P.I.W., Maller, J., Yelensky, R., Altshuler, D., and Daly, M.J. (2006). Evaluating and improving power in whole-genome association studies using fixed marker sets. Nat. Genet. *38*, 663–667.

53. Bhangale, T.R., Rieder, M.J., and Nickerson, D.A. (2008). Estimating coverage and power for genetic association studies using near-complete variation data. Nat. Genet. *40*, 841–843.

54. Manolio, T.A., Brooks, L.D., and Collins, F.S. (2008). A HapMap harvest of insights into the genetics of common disease. J. Clin. Invest. *118*, 1590–1605.

55. Hirschhorn, J.N. (2009). Genomewide association studies--illuminating biologic pathways. N. Engl. J. Med. *360*, 1699–1701.

56. Lettre, G., and Rioux, J.D. (2008). Autoimmune diseases: insights from genome-wide association studies. Hum. Mol. Genet. *17*, R116–21.

57. Hirschhorn, J.N., and Lettre, G. (2009). Progress in genome-wide association studies of human height. Horm. Res. *71 Suppl 2*, 5–13.

58. Zelcer, N., Hong, C., Boyadjian, R., and Tontonoz, P. (2009). LXR regulates cholesterol uptake through Idol-dependent ubiquitination of the LDL receptor. Science *325*, 100–104.

59. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.

60. Iles, M.M. (2008). What can genome-wide association studies tell us about the genetics of common disease? PLoS Genet. *4*, e33.

61. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. U. S. A. *106*, 9362–9367.

62. McCarthy, M.I., and Hirschhorn, J.N. (2008). Genome-wide association studies: potential next steps on a genetic journey. Hum. Mol. Genet. *17*, R156–65.

63. Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H., and Nadeau, J.H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet. *11*, 446–450.

64. Cirulli, E.T., and Goldstein, D.B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat. Rev. Genet. *11*, 415–425.

65. Zeng, D., and Lin, D.Y. (2005). Estimating haplotype-disease associations with pooled genotype data. Genet. Epidemiol. *28*, 70–82.

66. Ito, T., Chiku, S., Inoue, E., Tomita, M., Morisaki, T., Morisaki, H., and Kamatani, N. (2003). Estimation of haplotype frequencies, linkage-disequilibrium measures, and combination of haplotype copies in each pool by use of pooled DNA data. Am. J. Hum. Genet. *72*, 384–398.

67. Shaw, S.H., Carrasquillo, M.M., Kashuk, C., Puffenberger, E.G., and Chakravarti, A. (1998). Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. Genome Res. *8*, 111–123.

68. Erlich, Y., Chang, K., Gordon, A., Ronen, R., Navon, O., Rooks, M., and Hannon, G.J. (2009). DNA Sudoku--harnessing high-throughput sequencing for multiplexed specimen analysis. Genome Res. *19*, 1243–1253.

69. Prabhu, S., and Pe'er, I. (2009). Overlapping pools for high-throughput targeted resequencing. Genome Res. *19*, 1254–1261.

70. Shental, N., Amir, A., and Zuk, O. (2010). Identification of rare alleles and their carriers using compressed se(que)nsing. Nucleic Acids Res. *38*, e179.

71. Raineri, E., Ferretti, L., Esteve-Codina, A., Nevado, B., Heath, S., and Pérez-Enciso, M. (2012). SNP calling by sequencing pooled samples. BMC Bioinformatics *13*, 239.

72. Futschik, A., and Schlötterer, C. (2010). The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. Genetics *186*, 207–218.

Chapter 2

Large-Scale Pooled Next-Generation Sequencing of 1077 genes to identify genetic causes of short stature

Sophie R. Wang[1,2,3]*, Heather Carmichael[4]*, Shayne F. Andrew[5], Timothy C. Miller[2], Jennifer E. Moon[2], Michael A. Derr[5], Vivian Hwa[5], Joel N. Hirschhorn[1,2,3], Andrew Dauber[2,3]

1. Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115
2. Division of Endocrinology, Boston Children's Hospital, Boston, Massachusetts 02115
3. Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts 02142
4. Harvard Medical School, Boston, Massachusetts 02115
5. Department of Pediatrics, Oregon Health and Science University, Portland, Oregon 97239

*These authors contributed equally.

Originally published as:
Wang SR*, Carmichael H*, et al. *Large-scale pooled next-generation sequencing of 1077 genes to identify genetic causes of short stature*. J Clin Endocrinol Metab, 2013. **98**(8): p. 1428-37.

ABSTRACT

The majority of patients presenting with short stature do not receive a definitive diagnosis. Advances in genetic sequencing allow for large-scale screening of candidate genes, potentially leading to genetic diagnoses. The purpose of this study was to discover genetic variants that contribute to short stature in a cohort of children with no known genetic etiology. A total of 192 children with short stature with no defined genetic etiology and 192 individuals of normal stature from the Framingham Heart Study were studied. Pooled targeted sequencing using next-generation DNA sequencing technology of the exons of 1077 candidate genes was performed. The numbers of rare nonsynonymous genetic variants found in case patients but not in control subjects, known pathogenic variants in case patients, and potentially pathogenic variants in IGF1R were determined. We identified 4928 genetic variants in 1077 genes that were present in case patients but not in control subjects. Of those, 1349 variants were novel (898 nonsynonymous). False-positive rates from pooled sequencing were 4% to 5%, and the false-negative rate was 0.1% in regions covered well by sequencing. We identified 3 individuals with known pathogenic variants in PTPN11 causing undiagnosed Noonan syndrome. There were 9 rare potentially nonsynonymous variants in IGF1R, one of which is a novel, probably pathogenic, frameshift mutation. A previously reported pathogenic variant in IGF1R was present in a control subject. In summary, large-scale sequencing efforts have the potential to rapidly identify genetic etiologies of short stature, but data interpretation is complex; Noonan syndrome may be an underdiagnosed cause of short stature.

INTRODUCTION

Growth is a fundamental biological process that occurs during childhood. With the exception of diabetes, short stature is one of the most common reasons for referral to a pediatric endocrinologist. In most cases, short stature is familial, consistent with a strong genetic influence on childhood and adult height. In some cases, however, children have severe short stature that is out of proportion to the parental heights or have short stature associated with syndromic features. Molecular defects associated with these rarer cases have, over the last decade, expanded the list of genes and biological pathways known to influence growth. Multiple mutations, for example, have been found in the GH pathway, not only in the GH gene (GH1) itself, but also downstream within the GHR (GH receptor), STAT5B, IGF1, IGFALS, and IGF1R (IGF-I receptor) genes[1]. Many genes underlying severe skeletal dysplasias associated with short stature have also been identified[2]. Despite these advances, the molecular causality in the vast majority of patients, including those with severe or syndromic short stature, remains unresolved. Consequently, most affected patients continue to be classified as having idiopathic short stature.

Genome-wide association (GWA) studies have enabled the identification of common genetic variants (frequencies >5%) influencing quantitative traits such as height. Indeed, recent GWA studies identified 180 genetic loci with common DNA sequence variants that influence human stature[3]. Intriguingly, these common variants are often located in or near genes that underlie syndromes of abnormal skeletal growth. This overlap suggests that rare variants in other genes highlighted by the GWA studies could have significant effects on growth.

To explore the role of rare genetic variants in short stature, we developed and applied large-scale candidate gene sequencing technologies[4] in a cohort of children with short stature of unknown cause. The selected list of 1077 candidates is composed of genes from identified GWA loci, genes known to cause syndromic short stature, and genes known to be involved in growth

plate biology or growth plate signaling. Herein, we report our initial screening and assessment of pooled exonic sequencing on DNA samples from 192 children with short stature and 192 control children of normal stature. We identified a large number of nonsynonymous variants present in case patients but not in control subjects. There are a number of possible analytical approaches to explore these data. First, one can search for variants that have previously been reported to be pathogenic. Second, one can search for novel variants within genes known to cause short stature and then perform further familial segregation and functional studies to validate those variants. Third, one can search for multiple likely deleterious variants in novel genes not previously known to cause short stature. In the current article, we discuss the first approach, looking for known pathogenic variants, and provide a more detailed analysis of rare genetic variants identified in IGF1R as an example of the second approach. Haploinsufficiency of IGF1R is known to cause significant short stature, and our data demonstrate the utility of large-scale sequencing and the critical need for careful interpretation of the resulting data. Future work will explore the other analytical approaches.

RESULTS

Description of cohort

Participants in this study included 192 subjects (106 male and 86 female), 75% of whom were white. The height z scores ranged from −2.05 to −7.01 SD (Figure 2.1). The ages of these subjects ranged from 3 to 22 years with a mean of 10.3 years. Seventy subjects (36.4%) had begun GH therapy for short stature before enrollment in the study. However, GH deficiency was diagnosed in only 31 subjects (16%); of these, 22 had isolated GH deficiency without additional pituitary hormone defects. For those subjects receiving GH therapy, only height z scores before

initiation of therapy are shown (Figure 2.1). An additional 14 subjects were thought to have known genetic syndromes, but clinical diagnostic testing for the suspected syndromes had not identified pathogenic variants. Twenty-nine subjects were reported to have developmental delay.

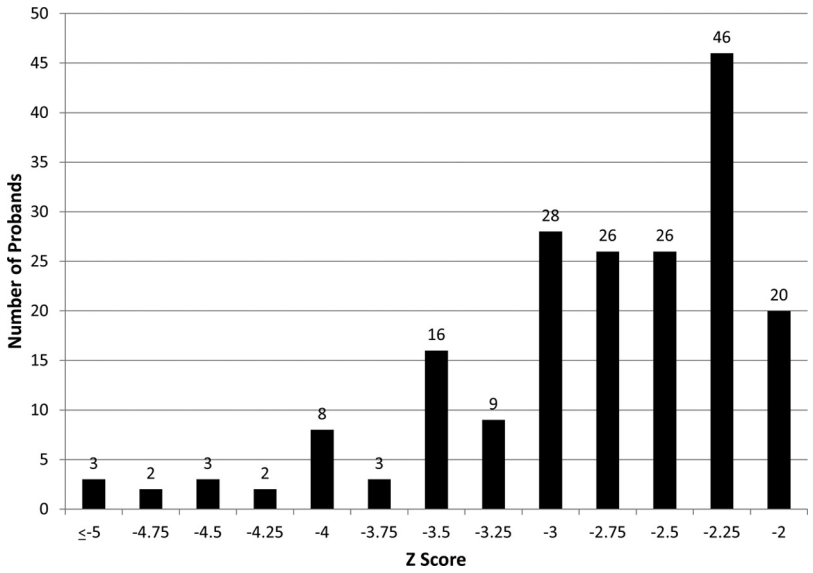

Figure 2.1: Height z score at enrollment or before initiation of GH therapy. Each bar represents individuals with a height z score less than or equal to the number noted below it on the x-axis but greater than the number below the bar to the left. For example, the right-most bar represents individuals with a height z score $-2.25 < z \leq -2$.

Validating pooled sequencing results

Using our pooled sequencing design, the false-positive rate of singleton variants estimated by comparison to the exome data was 4.8% (range, 0%–12.5% per individual) (Table 2.1). In addition, a total of 7 singleton variants mapped to the 8 holes in the 2 matrices compared with 7680 singleton variants that mapped to the 384 subjects (patients and control subjects), resulting in a similar estimation of the false-positive rate of 4.2%. These numbers establish the upper bound for the variant false-positive rate, because singleton variants are more likely to be false positives than variants found in multiple individuals. Of a total of 6618 variants present in the 6

exome samples within our target region, 7 variants were not identified by the pooled sequencing, giving an estimated overall false-negative rate of 0.1%. Similar to the false-positive rate, the false-negative rate for singleton variants is likely to be higher than that of other variants because singleton variants only appear in 2 pools and are more difficult to identify.

Table 2.1: False-positive rate of singletons estimated by comparing with exome sequencing of 6 samples

Sample No.	No. of singletons	False-positive singletons	False-positive rate, %
1	34	4	11.8
2	58	1	1.7
3	55	3	5.5
4	15	0	0.0
5	30	0	0.0
6	16	2	12.5
Summary	208	10	4.8

Pathogenicity of rare variants

In the 192 short stature patients, we identified a total of 10819 variants, of which 4928 were not detected in the control samples. Of these, 1349 were novel (Table 2.2). To screen for possible causal effects of variants found in our cohort, we compared these variants to those found in the Human Gene Mutation Database (HGMD)[5]. The database contains 26995 SNPs or indels located in the 1077 genes in our study, in which the variant has been reported as being associated with a particular clinical phenotype. We identified 66 such SNPs that matched a variant detected in our case subjects but not in control subjects. Because the HGMD is known to have erroneous entries,

we eliminated 7 variants with a minor allele frequency of ≥1% because these are unlikely to be true pathogenic variants. Of the remaining 59 variants, 32 were associated with recessive conditions or predispositions to complex traits, and the clinical pictures of the patients were not consistent with the disease phenotype, suggesting that they are unaffected heterozygous carriers. The final 27 variants previously associated with dominantly inherited diseases are listed in Table S2.1. We reviewed the phenotypes of the 27 case subjects and identified 1 case of autosomal dominant brachyolmia type 3 and 3 cases of Noonan syndrome. The remaining 24 case subjects did not have phenotypes consistent with the reported disease associations.

Table 2.2: Variants identified in short stature samples but not in control samples

Variant type	Known variants			Novel variants
	ALL	MAF≤5%	MAF≤1%	
Silent	1632	1602	1356	451
Missense	1903	1888	1704	829
Splice	11	11	10	8
Indel	11	11	10	46
Nonsense	22	22	22	15
Total	3579	3534	3102	1349

Abbreviation: MAF, minor allele frequency.

Identification of pathological variants associated with brachyolmia and Noonan syndrome

The patient with brachyolmia has a height of −3.88 SD and platyspondyly of the cervical spine. The mutation in TRPV4 is a missense mutation (c.1858G>A, V620I) that is a known variant causing the disease[6]. Before the research results became available but subsequent to

enrollment in our study, brachyolmia was clinically diagnosed in the patient, and clinical testing revealed this mutation.

All 3 patients with Noonan syndrome carried variants in PTPN11, the most common causative gene in this syndrome. Noonan syndrome is an autosomal dominant condition with characteristic dysmorphic facial features as well as short stature, webbed neck, and cardiac abnormalities[7]. The first subject is an 11-year-old girl with a height z score of −2.7 SD. Isolated GH deficiency was diagnosed at age 7 years, and she had a poor response to GH therapy. Of note, she was born with a transitional atrioventricular canal defect, which was repaired at 4 months of age. She had a triangular face with a mildly low posterior hair line and slightly wide-spaced eyes. She did not have ptosis or downslanting eyes, and her ears were normal. She carries the c.188A>G/p.Y63C variant[8]. The second subject is an 8-year-old girl with a height z score of −1.7 SD. She reached a height nadir of −3.0 SD at age 5 before the start of GH therapy for an indication of being small for gestational age to which she had a good response. She was evaluated at age 4 by a geneticist for the possibility of Russell-Silver syndrome, but no formal diagnosis was made. She has ptosis, epicanthal folds, downslanting eyes, and posteriorly rotated ears. She carries the c.925A>G/p.I309V variant[9], which she inherited from her father whose height is 173 cm (−0.5 SD). The third subject is a 16-year-old male adolescent with a height z score of −2.5 SD. He reached a height nadir of −3.2 SD at age 13 before isolated GH deficiency was diagnosed and GH therapy was started. He also started testosterone therapy at age 15 years for delayed puberty. He has mild learning issues, and on examination has a low posterior hair line but no other facial features consistent with Noonan syndrome. He carries the c.853T>C/p.F285L variant[9].

Identification of one pathological IGF1R variant among all IGF1R rare variants identified

To demonstrate the utility of a large-scale sequencing approach and the need for careful interpretation of results, we focused on rare variants in IGF1R, a gene for which haploinsufficiency is known to cause significant short stature[10–12]. Our approach was to identify nonsynonymous variants present in case patients only that segregated with the phenotype of short stature within the families. Variants meeting these criteria would be classified as potentially pathogenic variants requiring further functional validation. In total, our targeted sequencing found 25 unique IGF1R variants in both case patients and control subjects. Of these, 16 were synonymous SNPs, most of which were common (minor allele frequency >0.01); these were not evaluated further. The remaining 9 variants included 6 missense, 1 frameshift, and 2 intronic variants. The intronic variants were found within 5 bp of an exon, which suggests a potential involvement in splicing, and thus were included for further analysis. Five of these variants were present in case patients only (Table 2.3). All 9 variants were validated via traditional Sanger sequencing and confirmed to be heterozygous. To determine the biological significance of these variants, segregation of variants 2 through 6 within families was performed (Figure 2.2). There was no correlation between the individual family member's heights and the carrier status of the variants, suggesting that these variants are not likely to be major contributors to the patients' short stature, and, therefore, we excluded these variants from further consideration as pathogenic variants. Variant 7 was present in multiple case patients and control subjects and was also not likely to be pathogenic. Of note, 1 of the 2 rare missense variants found only in control subjects in our study (variant 9) was previously reported as pathogenic in the literature[13]. This control subject is of normal stature at −0.4 SD.

Table 2.3: IGF1R Potentially Nonsynonymous Variants

Variant	Exon	cDNA	Protein	MAF	Subject	Sex	Height SD	Birth weight, g[a]	IGF-1 (normal range[b]), ng/mL
1	2	c.418dupG	p.A140Gfs*5	novel	Case 1	F	-4.1	unknown	389.8 (244 - 787)
2	5	c.1247+3A>G	Intron	0.0007	Case 2	F	-3.9	2800	26.6 (49 - 342)
3	7	c.1463-5C>A	Intron	0.012	Case 3	M	-2.4	3500	52.2 (49 - 342)
4	6	c.1411C>T	p.R471C	novel	Case 4	M	-2.3	4100	34 (63 - 279)
5	7	c.1502C>T	p.S501L	novel	Case 5	M	-3.0	3400	148 (63 - 279)
6	6	c.1336A>G	p.M446V	0.0027	Case 6	F	-2.8	2600	97.2 (49 - 342)
					Control 6	F	0.0		

Table 2.3 (Continued)

7	6	c.1310G>A	p.R437H	0.004	Case 7a	F	-3.3	4200	69.9 (49 - 342)
					Case 7b	F	-3.1	3800	62 (63 - 279)
					Control 7a	F	+0.4		
					Control 7b	M	0.0		
					Control 7c	M	-0.4		
8	5	c.1162G>A	p.V388M	0.003	Control 8	M	-0.6		
9	7	c.1532G>A	p.R511Q	0.003	Control 9	M	-0.4		

Abbreviations: F, female; M, male.

a. All case patients were the product of full-term pregnancies (>37 weeks) with the exception of case 7b (36.5 weeks). None of the case patients met the definition for intrauterine growth retardation (weight <2500 g at birth for normal gestation).

b. Normal range for sex and Tanner stage. All IGF-I values were obtained during a baseline clinical evaluation and were measured when the patient was not receiving growth hormone therapy.

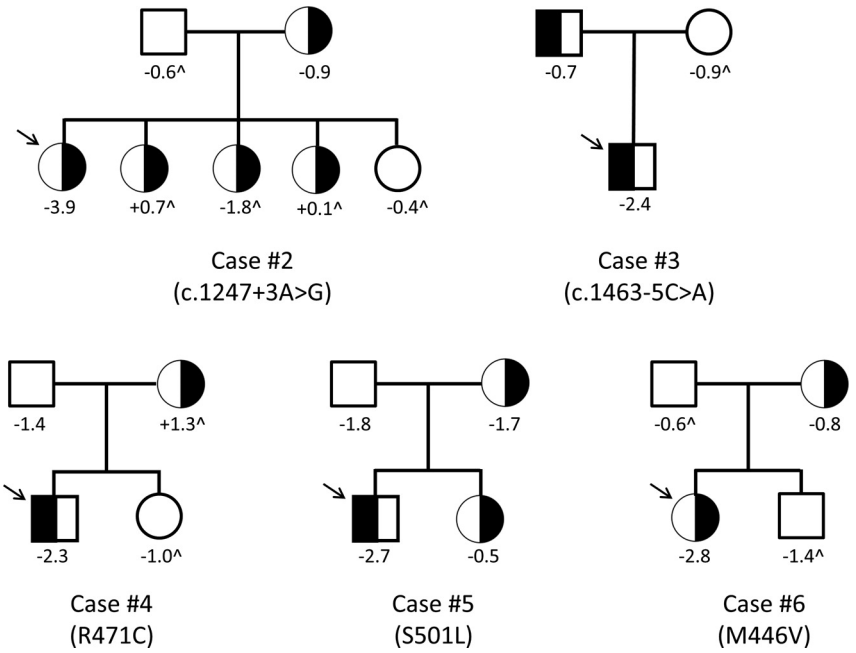

Figure 2.2: Segregation of identified IGF1R nonsynonymous variants in affected families does not correlate with short stature. Numbers below the individuals denote the height z scores. ^ indicates that the height was estimated by a family member. All other values were measured. Individuals carrying the heterozygous variants are indicated as black half-filled circles (females) or squares (males). The arrow points to the affected proband in each family.

Variant 1 was a novel frameshift mutation (c.418dupG/p.A140Gfs*5) (Figure S2.1) found in 1 patient in the heterozygous state. The mutation causes severe truncation of the protein with complete abrogation of the transmembrane and intracellular domains and thus was predicted to lead to haploinsufficiency. This patient was adopted from China at 6 years of age, and therefore a complete history and familial samples could not be obtained. At the age of 15 years, the patient had Tanner stage 4 breast development with height of 136 cm (−4.06 SD), weight of 30.2 kg (−4.87 SD), and a head circumference of 49.3 cm (−4.4 SD). She has a history notable for bilateral cleft lip and palate as well as

attention deficit disorder and mild developmental delay. Her IGF-I was normal at 389.8 ng/mL (normal range, 244–787 ng/mL for a Tanner stage 4 female). GH stimulation testing with arginine and glucagon demonstrated a peak GH level of 18.8 ng/mL. She had previously been treated with GH therapy with a possible mild increase in growth velocity, although this occurred concurrently with entering puberty.

Variant 1 was the only variant in IGF1R that met our prespecified criteria for consideration as a potential pathogenic variant. To determine whether variant 1 was causal for the patient's phenotype, we evaluated IGF1R expression and function in primary PBMCs derived from the patient compared with those in control PBMCs (procured from the unrelated adoptive mother). Flow cytometric analysis by FACS of live PBMCs (counts, y-axis; Figure 2.3) indicated that fluorescence emitted by IGF1R-PE-labeled PBMCs was markedly reduced (fluorescence intensity, x-axis; Figure 2.3) in patient PMBCs compared with that by the normal control PBMCs (Figure 2.3A). When the live PBMCs were treated with IGF-I, emitted fluorescence was comparably reduced for both control and patient PBMCs, suggesting normal internalization of IGF1R upon ligand binding (Figure 2.3B). Immunoblot analysis of cell lysates, furthermore, indicated that total IGF1R expression was reduced in the patient's PBMCs with correlating reductions in IGF-I–induced signaling (Figure 2.3C). Taken together, the results are consistent with the heterozygous IGF1R c.418dupG variant inducing a state of IGF1R deficiency and being an excellent candidate to cause the subject's short stature.

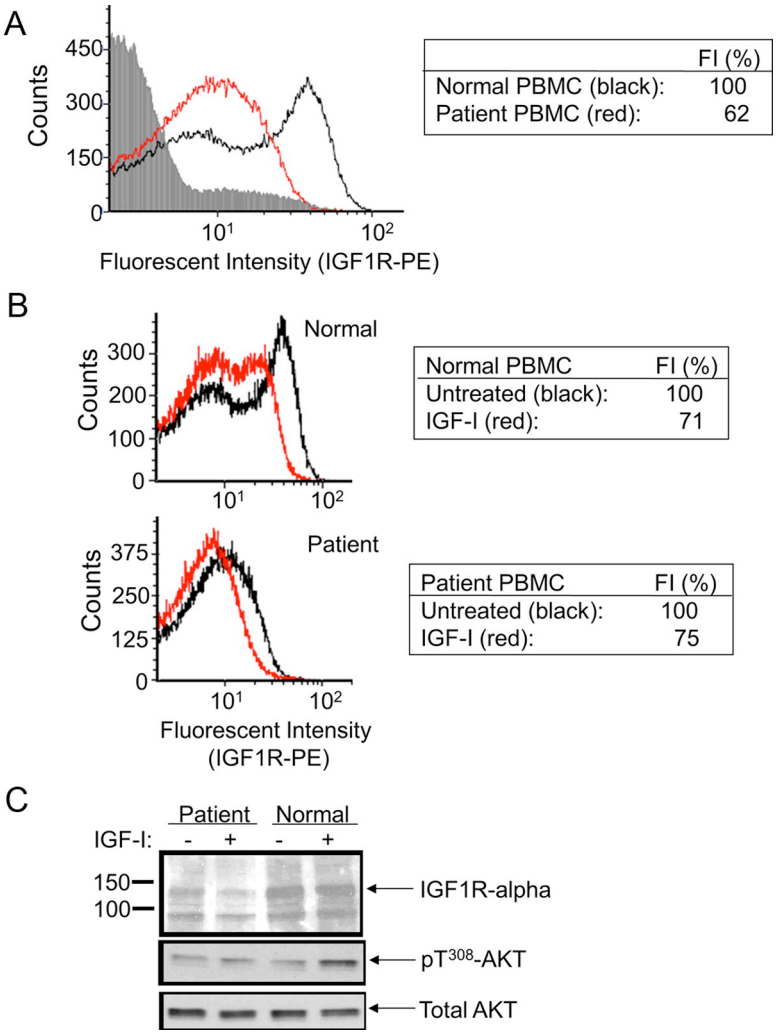

Figure 2.3: IGF1R expression and signaling in primary peripheral PBMCs of patient carrying heterozygous IGF1R c.418dupG. PBMCs were isolated as indicated in Materials and Methods. Flow cytometry analysis by FACS was used to detect IGF1R, labeled by PE-conjugated anti-human IGF1R-α antibody (see Materials and Methods), on the cell surface of live PBMCs. Live PBMCs (counts, y-axis) and fluorescence emitted by the IGF1R-PE–labeled PBMC were collated (log scale fluorescence intensity, x-axis). (A) Patient (red graph) compared with normal (black graph) PBMCs. Background fluorescence emitted by unlabeled and untreated PBMC control is shown by the gray-shaded region. The geometric mean of the fluorescent intensity (FI) detected in normal PBMCs was given an arbitrary unit of 100% (table). (B) Effect of IGF-I treatment (100 ng/mL, 1 hour) on the detection of IGF1R-PE–labeled PBMC from normal (top panel) and patient (bottom panel) PBMCs. For each, the geometric mean of the fluorescent intensity (FI) detected in untreated PBMCs was given an arbitrary unit of 100%. (C) Western immunoblot analysis of total cell lysates from PBMCs treated with IGF-I (100 ng/mL) vs untreated cells. Molecular mass (kilodaltons) is indicated on the left side of the immunoblots. The intracellular proteins detected are indicated by arrows (on right).

DISCUSSION

Short stature is a common problem confronting pediatric endocrinologists. After exclusion of other chronic diseases or overt hormonal deficiencies, clinicians are often unable to provide a definitive diagnosis for the etiology of an individual patient's short stature. There are a multitude of genetic causes for short stature, but most patients do not fall into a previously identified genetic syndrome. We, therefore, designed and performed a large-scale sequencing project to identify pathogenic rare genetic variants in individuals with short stature. We sequenced 1077 candidate genes including known skeletal dysplasia genes, genes within the GH signaling pathway, genes known to affect growth plate biology, and genes within loci associated with adult height in large GWA studies. Using this approach, we identified 4 known pathogenic variants causing short stature as well as novel variants in genes known to affect stature.

To facilitate the sequencing of a large number of genes in many subjects, we used a pooled sequencing design, which significantly reduced the cost of such analysis[14]. Most of the cost associated with a targeted next-generation sequencing project is typically incurred at the library construction stage, in which targeted regions of DNA are separated from the remainder of the genome for sequencing. In a pooled sequencing design, this process only has to be performed once per pool. Although actual sequencing depth may need to be increased to ensure adequate representation of all samples in the pool, the associated cost is relatively minor. Indeed, the cost per sample of our pooled targeted sequencing approach is estimated to be ~15% of the cost for individual exome sequencing (ie, ~$110 compared with ~$800 per sample, based on current prices available at our institution). Exome sequencing could also be done in a pooled fashion, in

which case cost differences will depend on the cost of sequencing coverage, a process that is becoming cheaper to perform. Although pooled exome sequencing does have the advantage that nearly all genes are evaluated, analysis and interpretation of the data generated would be much more complex because of the large number of novel nonsynonymous variants in genes with no known connection to the phenotype of interest. Our simple matrix pooling design, in contrast, allows for the rapid assessment of low-frequency variants in candidate genes and the identification of individuals carrying singleton variants, which are more likely to be pathogenic than variants with a higher minor allele frequency. However, pooling does limit the ability to discern whether a single variant is homozygous or heterozygous in an individual subject and follow-up confirmatory genotyping is necessary. Using this design, we were able to identify a large number of very rare nonsynonymous variants within our candidate genes with low false -positive and low false-negative rates.

We identified 4 subjects in our cohort who had known pathogenic variants implicated in disease. Notably, 3 of these subjects have mutations in PTPN11 that cause Noonan syndrome. Noonan syndrome is known to have a wide phenotypic spectrum, leading to difficulty in diagnosis[7], and, indeed, one of our subject's fathers carries a proven pathogenic variant in PTPN11 yet never presented with the overt clinical manifestations of Noonan syndrome. Although it is true that our subjects may have had features consistent with Noonan syndrome that were unrecognized, such as a cardiac defect or delayed puberty, this retrospective recognition of related features does not eliminate the benefit of genetic screening. The lack of diagnoses in our cohort represents clinical reality, because these subjects were extensively evaluated by experienced

pediatric endocrinologists and in one case by geneticist as well. This suggests that a substantial number of patients with Noonan syndrome are designated as having idiopathic short stature or isolated GH deficiency even after clinical evaluation by pediatric subspecialists. Additional research is needed to determine whether widespread screening for PTPN11 or the other Noonan syndrome genes is warranted for patients with short stature of unknown etiology.

It is important to note that, in our cohort, the vast majority of HGMD-reported disease-causing dominant mutations did not manifest with the associated clinical phenotype. The classification of these variants as pathogenic is probably erroneous and based on insufficient clinical evidence. However, we cannot rule out the possibility that the variants have variable expressivity and some of our subjects are presenting on the very mild end of the clinical spectrum with short stature as their disease manifestation. Our focus on rare variants of the IGF1R gene illustrates the critical importance of providing supporting familial segregation and functional data when a rare variant has been identified. IGF-I, the primary mediator of GH function, is essential for growth. Heterozygous and compound heterozygous mutations in IGF1R that lead to decreases in the quantity or function of the receptor have been described in nearly a dozen human cases[1,10–12,15,16]. These patients display variable phenotypes, with shared characteristics that include poor prenatal and postnatal growth, microcephaly, high or normal IGF-I levels, and developmental delay[1,10–12,15,16].

Our targeted sequencing approach identified 7 unique rare nonsynonymous IGF1R variants as well as 2 intronic variants with the potential to affect splicing because of their proximity to exons. Only 1 of these 9 variants, a novel

c.418dupG frameshift mutation located in exon 2, was associated with clinical features suggestive of a pathological IGF1R deficiency state (high levels of serum IGF-1, microcephaly, and intrauterine growth retardation). Furthermore, in primary cells derived from this patient, significant decreases in both IGF1R expression and IGF-I–induced signaling supported the pathogenicity of the IGF1R c.418dupG defect. None of the remaining variants found in case patients show convincing evidence of pathogenicity. This example demonstrates that large-scale sequencing efforts will identify numerous very rare and novel nonsynonymous variants in candidate genes. Most of these variants will be missense variants, leading to a change in a single amino acid, which will not affect protein function, and represent incidental findings. Segregation of these variants with the phenotype within families is the first critical step in evaluating potential pathogenicity, highlighting the importance of collecting familial samples at the time of the initial DNA collection.

Filtering strategies based on population allele frequency are useful and necessary, but most public databases do not provide individual phenotypic data linked to the subject's genotype, thus limiting the ability to determine whether a variant is potentially pathogenic. Therefore, simultaneous sequencing of a control cohort with a known phenotype, in this case normal stature, provides additional information about the lack of pathogenicity of rare variants in a gene. This fact is exemplified by our finding that an IGF1R variant previously reported to be pathogenic (c.1532G>A/p.R511Q)[13] was found in the heterozygous state in a control subject of normal stature. This variant was originally identified in the heterozygous state in a patient and her maternal aunt, both of whom presented with extreme short stature (−6.1 and −5.7 SD, respectively). It is of note

that information regarding the parents of the patient were lacking in this report. In vitro reconstitution studies of the homozygous p.R511Q variant were performed to support the pathogenicity of this variant, although the effect of heterozygosity was unknown. These caveats, together with our identification of the same variant in a control subject of normal stature, strongly suggest that a heterozygous p.R511Q is not likely to be the cause of the previously reported family's extreme short stature. Furthermore, Kansra et al[17] recently detected the R511Q variant in 6 of 1800 public school students. Indeed, carriers had an average height around the 27th percentile, thus providing additional evidence that this variant does not cause severe short stature. Taken together, these results support the importance of segregation analysis and the need to include primary cells in functional analysis.

Our study has a number of important limitations. We recruited a very heterogeneous cohort, allowing for the inclusion of dysmorphic features, other congenital anomalies, and hormonal deficiencies provided that there was no known genetic etiology for these findings. Thus, subjects in this cohort do not meet a strict definition of idiopathic short stature[18]. Nevertheless, we believe that this cohort more accurately represents the diversity of patients who are seen in a referral setting and is probably enriched for individuals with rare genetic variants that may have multisystem effects. In addition, our hybrid selection strategy only targets the exons of the candidate genes, and, thus, any noncoding variation that affects gene expression cannot be detected by our methods. Variants affecting gene expression can play an important role in causing short stature. For example, Russell-Silver syndrome, an important syndromic form of short stature, is often due to abnormalities in methylation of chromosome 11p15.5, leading to aberrant gene

expression[19]. In addition, our current approach does not assess copy number variation (ie, deletions or duplications of genes), which may also be an important genetic defect leading to short stature. We are currently pursuing copy number analysis of this cohort using a custom chromosomal microarray (data not shown). Furthermore, we did not obtain perfect sequencing coverage of all variants in the targeted region and could miss potentially pathogenic variants in the candidate genes. Finally, because of the large numbers of rare missense variants in both case patients and control subjects, we have limited power to discover new genes with a statistically significant excess of mutations in case patients vs control subjects. Ongoing work to increase sample size and examine subjects at the extremes of the height distribution will provide additional data to support novel gene discovery.

In conclusion, we present the initial results of a large-scale candidate gene sequencing effort in children with short stature and demonstrate the complexity of data interpretation of such efforts. Of our 192 subjects, 3 were found to have known pathogenic variants in PTPN11, highlighting the possibility that Noonan syndrome is underdiagnosed in the clinical setting. We report a novel frameshift mutation in IGF1R and demonstrate its pathogenicity in vivo. In addition, we provide evidence that a previously reported variant in IGF1R is not pathogenic. Analyses of variants identified in the other candidate genes are currently ongoing.

MATERIALS AND METHODS

Height candidate genes

In this study, we sequenced the exons of 1077 genes (~2 Mb total target size). Of these 1077 genes, one-third (n = 356) were known biological candidates, including genes known to underlie syndromic growth disorders or skeletal dysplasias as well as genes involved in growth plate biology or GH signaling. The remaining two-thirds (n = 777) included genes within genomic loci associated with height based on GWA studies; 56 genes belong to both categories (see Table S2.2)[3]. For the genes within the GWA loci, we set the genomic boundaries at each height-associated locus using linkage disequilibrium cutoffs (HapMap CEU r2 > 0.5) for the top single-nucleotide polymorphism (SNP). For loci with ≥2 genes within the genomic boundary, all genes were included. Loci with >10 genes were excluded. For SNPs with <2 genes within the genomic boundary, genes beyond the boundary but within the next recombination hotspots were included.

Subjects

This study was approved by the institutional review board at Boston Children's Hospital (Boston, Massachusetts). All subjects or their legal guardians provided written informed consent. The 192 patients with short stature (>2 SD below the mean for age and sex)[20] but without defined genetic etiologies, were recruited from the endocrinology and genetics clinics at Boston Children's Hospital. Because we were searching for rare genetic syndromes, subjects were allowed to have additional medical comorbidities, dysmorphic features, or other hormonal deficiencies as long as these alternate medical problems did not provide a clear explanation for the subject's short stature. In addition, 192 control subjects were chosen from the Framingham Heart Study. Control subjects were chosen from the middle of the Framingham Heart Study height distribution (height

z scores between −0.7 and +0.7 SDs). z scores were calculated by regressing the height phenotype, stratifying by sex and adjusting for age.

Sequencing protocol

DNA samples from multiple subjects were pooled for DNA sequencing using previously described methods available at the Broad Institute[21]. To identify variants present only in a single individual (hereon referred to as singleton variants), we applied a simple overlapping pooling design (Figure S2.2). The samples from short stature subjects and control samples were each arranged into a 14 × 14 matrix of 28 pools, with 13 to 14 samples in each pool. Four empty "holes" were included in each matrix for assessing the false-positive rate. Each sample was sequenced in 2 pools (1 row pool and 1 column pool). Singleton variants appear only in 1 row pool and 1 column pool. Therefore, the subject whose DNA sample is present at the intersection of these 2 pools must be the individual carrying that singleton variant. The targeted exons of the 1077 candidate genes were enriched using a custom Agilent SureSelect hybrid selection system. Sequencing was performed on the Illumina HiSeq platform. There was an average of 12961604 reads per pool, resulting in mean target coverage of 213 reads (15 reads per subject in a pool of 14 subjects or 30 total reads per subject, as each subject is present in 2 pools). Variant calling was performed using Syzygy software[21] and then a new likelihood-based secondary calling strategy that integrated the extra information from our matrix design was applied (Supplemental Methods).

Variants were annotated for functional effect using SnpEff 2.0.5 (http://snpeff.sourceforge.net/). Variant allele frequency data were obtained from 3

publicly available datasets: (1) the integrated variant call set of 1000 Genomes phase 1 samples[22] (February 2012 release); (2) the National Heart, Lung, and Blood Institute Exome Variant Server[23], and (3) ~12 000 sequenced genomes and exomes assembled for exome genotyping chip design (http://genome.sph.umich.edu/wiki/Exome_Chip_Design). The maximal allele frequencies from all 3 sources were used. Novel variants are those not observed in any of these datasets.

Assessing false-positive and false-negative rates

We estimated the false-negative and false-positive rates by comparing pooling data with data from exome sequencing previously performed in 6 of the short stature subjects. To start, we determined the overlapping targets between pooling and exome capture arrays. Then, limiting to sites with ≥10 reads, we assumed that the exome sequencing data reflected the gold standard because of its much greater depth of coverage. False positives were defined as singletons observed in pooling data but not in exome data, whereas false-negative results were those observed in exome data but not in either of the 2 relevant pools. The false-positive rate was also estimated by looking for singleton variants that mapped to one of the empty holes in the matrix. The singleton variants that mapped to empty holes were false-positives, permitting the number of false-positive variants per individual to be estimated, and from there we could independently estimate the false-positive rate of singleton variants.

IGF1R functional studies

Whole blood samples (BD Vacutainer Cell Preparation Tube with sodium heparin; Becton, Dickinson and Company, Franklin Lakes, New Jersey) were collected from the adopted patient and from the unrelated mother, who served as normal control subjects. Peripheral blood mononuclear cells (PBMCs) were isolated following the manufacturer's protocol. PBMCs, in freezing medium (RPMI 1640 medium + 40% fetal bovine serum + 10% dimethyl sulfoxide) were stored in liquid nitrogen.

For immunoblot analysis, fresh PBMCs (2 × 106/treatment) were resuspended in serum-free RPMI 1640 medium, with or without recombinant IGF-I (100 ng/mL; GroPep Ltd, Thebarton, South Australia, Australia), for 20 minutes at 37°C in a CO2 incubator, before pelleting, and cells were lysed as described previously for fibroblast cell cultures[24]. Western immunoblot analyses were performed as described previously[24]. For flow cytometry analysis by fluorescence-activated cell sorting (FACS) of cell surface IGF1R, PBMCs, warmed to 37°C from liquid nitrogen storage, were washed twice, aliquoted as 1 × 106cells/sample in RPMI 1640 medium + 10% fetal bovine serum, and incubated overnight at 37°C (5% CO2 incubator). Before IGF-I treatment, cells were washed twice with serum-free RPMI 1640 medium + 0.5% BSA and equilibrated in 0.5 mL of serum-free RPMI 1640 medium for 4 hours. Cells were treated with or without IGF-I (100 ng/mL final concentration) for 1 hour, after which cells were washed twice with cold staining medium (1× PBS-0.5% BSA-0.1% sodium azide) and incubated with phycoerythrin (PE)–conjugated anti-human IGFIR-α (CD221; BD Biosciences, San Jose, California) for 30 minutes at 4°C in the dark. After antibody staining, cells were washed twice with cold staining medium, resuspended in 200 μL of staining medium-0.25% propidium iodide, and incubated on ice for 10 minutes. A total of 100 000 live PBMCs

(propidium iodide negative, CD221 positive) per sample were acquired via a FACSCaliber flow cytometer (BD Biosciences), and the fluorescence emitted by IGF1R-PE–labeled PBMCs was analyzed using FCS Express 3 analysis software (De Novo Software, Los Angeles, California).

ACKNOWLEDGEMENT

We thank Jason Flannick for his assistance in running the Syzygy software and Dr. Amy Roberts for her helpful discussions regarding Noonan syndrome. This work was supported by Harvard Catalyst, The Harvard Clinical and Translational Science Center (National Institutes of Health [NIH] Award UL1 RR 025758 and financial contributions from Harvard University and its affiliated academic health care centers). The content is solely the responsibility of the authors and does not necessarily represent the official views of Harvard Catalyst, Harvard University and its affiliated academic health care centers, the National Center for Research Resources, or the NIH. Sequencing experiments were performed by the Sequencing Core Facility of the Molecular Genetics Core Facility at Boston Children's Hospital supported by NIH P30-HD18655. This work was also supported by NIH Grant 1K23HD073351 (to A.D.), a fellowship grant from the Genentech Center for Clinical Research in Endocrinology (to A.D.), the Translational Research Program at Boston Children's Hospital, and March of Dimes Grant 6-FY09-507 (to J.N.H.). Samples were provided from the Framingham Heart Study of the National Heart, Lung, and Blood Institute of the NIH and Boston University School of Medicine, which was supported by the National Heart, Lung, and Blood Institute Framingham Heart Study (Contract N01-HC-25195).

NOTE ADDED AT PROOF

After this paper was published, we were informed that PTPN11 p.I309V exists at 7% in Ashkenazi Jews. Thus this variant is likely benign.

REFERENCES

1. David, A., Hwa, V., Metherell, L.A., Netchine, I., Camacho-Hübner, C., Clark, A.J.L., Rosenfeld, R.G., and Savage, M.O. (2011). Evidence for a continuum of genetic, phenotypic, and biochemical abnormalities in children with growth hormone insensitivity. Endocr. Rev. *32*, 472–497.

2. Rimoin, D.L., Cohn, D.H., and Eyre, D. (1994). Clinical--molecular correlations in the skeletal dysplasias. Pediatr. Radiol. *24*, 425–426.

3. Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature *467*, 832–838.

4. DePristo, M.A., Banks, E., Poplin, R., Garimella, K. V, Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

5. Stenson, P.D., Ball, E., Howells, K., Phillips, A., Mort, M., and Cooper, D.N. (2008). Human Gene Mutation Database: towards a comprehensive central mutation database. J. Med. Genet. *45*, 124–126.

6. Rock, M.J., Prenen, J., Funari, V.A., Funari, T.L., Merriman, B., Nelson, S.F., Lachman, R.S., Wilcox, W.R., Reyno, S., Quadrelli, R., et al. (2008). Gain-of-function mutations in TRPV4 cause autosomal dominant brachyolmia. Nat. Genet. *40*, 999–1003.

7. Romano, A.A., Allanson, J.E., Dahlgren, J., Gelb, B.D., Hall, B., Pierpont, M.E., Roberts, A.E., Robinson, W., Takemoto, C.M., and Noonan, J.A. (2010). Noonan syndrome: clinical features, diagnosis, and management guidelines. Pediatrics *126*, 746–759.

8. Tartaglia, M., Mehler, E.L., Goldberg, R., Zampino, G., Brunner, H.G., Kremer, H., van der Burgt, I., Crosby, A.H., Ion, A., Jeffery, S., et al. (2001). Mutations in PTPN11,

encoding the protein tyrosine phosphatase SHP-2, cause Noonan syndrome. Nat. Genet. *29*, 465–468.

9. Tartaglia, M., Kalidas, K., Shaw, A., Song, X., Musat, D.L., van der Burgt, I., Brunner, H.G., Bertola, D.R., Crosby, A., Ion, A., et al. (2002). PTPN11 mutations in Noonan syndrome: molecular spectrum, genotype-phenotype correlation, and phenotypic heterogeneity. Am. J. Hum. Genet. *70*, 1555–1563.

10. Ester, W.A., van Duyvenvoorde, H.A., de Wit, C.C., Broekman, A.J., Ruivenkamp, C.A.L., Govaerts, L.C.P., Wit, J.M., Hokken-Koelega, A.C.S., and Losekoot, M. (2009). Two short children born small for gestational age with insulin-like growth factor 1 receptor haploinsufficiency illustrate the heterogeneity of its phenotype. J. Clin. Endocrinol. Metab. *94*, 4717–4727.

11. Kawashima, Y., Kanzaki, S., Yang, F., Kinoshita, T., Hanaki, K., Nagaishi, J.-I., Ohtsuka, Y., Hisatome, I., Ninomoya, H., Nanba, E., et al. (2005). Mutation at cleavage site of insulin-like growth factor receptor in a short-stature child born with intrauterine growth retardation. J. Clin. Endocrinol. Metab. *90*, 4679–4687.

12. Abuzzahab, M.J., Schneider, A., Goddard, A., Grigorescu, F., Lautier, C., Keller, E., Kiess, W., Klammt, J., Kratzsch, J., Osgood, D., et al. (2003). IGF-I receptor mutations resulting in intrauterine and postnatal growth retardation. N. Engl. J. Med. *349*, 2211–2222.

13. Inagaki, K., Tiulpakov, A., Rubtsov, P., Sverdlova, P., Peterkova, V., Yakar, S., Terekhov, S., and LeRoith, D. (2007). A familial insulin-like growth factor-I receptor mutant leads to short stature: clinical and biochemical characterization. J. Clin. Endocrinol. Metab. *92*, 1542–1548.

14. Golan, D., Erlich, Y., and Rosset, S. (2012). Weighted pooling--practical and cost-effective techniques for pooled high-throughput sequencing. Bioinformatics *28*, i197–206.

15. Fang, P., Cho, Y.H., Derr, M.A., Rosenfeld, R.G., Hwa, V., and Cowell, C.T. (2012). Severe short stature caused by novel compound heterozygous mutations of the insulin-like growth factor 1 receptor (IGF1R). J. Clin. Endocrinol. Metab. *97*, E243–7.

16. Walenkamp, M.J.E., van der Kamp, H.J., Pereira, A.M., Kant, S.G., van Duyvenvoorde, H.A., Kruithof, M.F., Breuning, M.H., Romijn, J.A., Karperien, M., and Wit, J.M. (2006). A variable degree of intrauterine and postnatal growth retardation in a family with a missense mutation in the insulin-like growth factor I receptor. J. Clin. Endocrinol. Metab. *91*, 3062–3070.

17. Kansra, A.R., Dolan, L.M., Martin, L.J., Deka, R., and Chernausek, S.D. (2012). IGF receptor gene variants in normal adolescents: effect on stature. Eur. J. Endocrinol. *167*, 777–781.

18. Cohen, P., Rogol, A.D., Deal, C.L., Saenger, P., Reiter, E.O., Ross, J.L., Chernausek, S.D., Savage, M.O., and Wit, J.M. (2008). Consensus statement on the diagnosis and treatment of children with idiopathic short stature: a summary of the Growth Hormone Research Society, the Lawson Wilkins Pediatric Endocrine Society, and the European Society for Paediatric Endocrinology Workshop. J. Clin. Endocrinol. Metab. *93*, 4210–4217.

19. Abu-Amero, S., Wakeling, E.L., Preece, M., Whittaker, J., Stanier, P., and Moore, G.E. (2010). Epigenetic signatures of Silver-Russell syndrome. J. Med. Genet. *47*, 150–154.

20. Kuczmarski, R.J., Ogden, C.L., Guo, S.S., Grummer-Strawn, L.M., Flegal, K.M., Mei, Z., Wei, R., Curtin, L.R., Roche, A.F., and Johnson, C.L. (2002). 2000 CDC Growth Charts for the United States: methods and development. Vital Health Stat. 11. 1–190.

21. Rivas, M.A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C.K., Boucher, G., Ripke, S., Ellinghaus, D., Burtt, N., et al. (2011). Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat. Genet. *43*, 1066–1073.

22. Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

23. National Heart, Lung, and B.I. Exome Variant Server.

24. Fang, P., Schwartz, I.D., Johnson, B.D., Derr, M.A., Roberts, C.T., Hwa, V., and Rosenfeld, R.G. (2009). Familial short stature caused by haploinsufficiency of the insulin-like growth factor i receptor due to nonsense-mediated messenger ribonucleic acid decay. J. Clin. Endocrinol. Metab. *94*, 1740–1747.

Chapter 3

Heterozygous mutations in natriuretic peptide receptor B (NPR2) gene as a cause of short stature

Sophie R. Wang[1,2,3*], Heather Carmichael[4*], Christina M. Jacobsen[5*], Timothy C. Miller[2], Jennifer E. Moon[2], Joel N. Hirschhorn[1,2,3], Andrew Dauber[2,3]

1. Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115
2. Division of Endocrinology, Boston Children's Hospital, Boston, Massachusetts 02115
3. Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts 02142
4. Harvard Medical School, Boston, Massachusetts 02115
5. Orthopaedic Research Laboratories, Department of Orthopaedic Surgery, Boston Children's Hospital, Boston, Massachusetts 02115

*These authors contributed equally.

ABSTRACT

Based on the observation of reduced stature in relatives of patients with acromesomelic dysplasia, type Maroteaux (AMDM), homozygous for mutations in natriuretic peptide receptor B gene (NPR2), it has been suggested that heterozygous mutations in this gene could be responsible for the growth impairment observed in some cases of idiopathic short stature (ISS), and some small subsequent studies provided support for this hypothesis. We enrolled 192 unrelated patients with short stature and 7 heterozygous NPR2 missense or loss-of-function mutations, including one *de novo* splice site variant, were identified. These allelic variants were not found in our controls and 5 of them were not found in the public databases. NPR2 mutations were also found in relatives, all of whom had short stature (height SDS below -2), consistent with a dominant inheritance pattern. We then went on to investigate the presence of NPR2 mutations in two larger cohorts of individuals selected from the extremes of height distribution. Functional studies of the NPR2 mutations identified are pending. With these functional results, we will be able to test more rigorously the hypothesis that NPR2 functional haploinsufficiency causes short stature.

INTRODUCTION

C-type natriuretic peptide (CNP) is a small, secreted peptide and a member of the natriuretic peptide family. CNP binds to a homodimeric transmembrane receptor, natriuretic peptide receptor B (NPR2), which functions as a guanylyl cyclase to generate cGMP in chondrocytes, female reproductive organs, and endothelial cells[1,2]. Further intracellular signaling occurs through cGMP-dependent protein kinases, cGMP binding phosphodiesterases, and cyclic nucleotide-gated ion channels.

Several lines of evidence indicate that CNP/NPR2 signaling is an important regulator of skeletal growth. CNP-overexpressing mice exhibit excessive growth[3], while defects of the CNP[4] or NPR2[5] gene, lead to impairment of skeletal development. In humans, loss-of-function mutations in NPR2 cause acromesomelic dysplasia, Maroteaux type (AMDM; OMIM 602875). This autosomal recessive skeletal dysplasia is characterized by dwarfism and short limbs[6]. On the other hand, overproduction of CNP due to a chromosomal translocation was reported to cause skeletal dysplasia associated with tall stature[7,8]. In addition, gain-of-function mutations of NPR2 have been identified in several studies to cause overgrowth disorder[9–11].

Interestingly, in the first report of biallelic NPR2 mutations causing AMDM, parents of patients with AMDM (obligate heterozygotes) were noted to be shorter than expected for their population of origin[6], though these individuals came from a wide range of geographic and ethnic backgrounds, and measurements were done by a number of observers, which complicated the comparison. Another study that evaluated a single family with an AMDM proband showed that the heterozygous carriers had a mean height 1.4 SD lower than their non-carrier family members[12]. In this single-family study, the proband's parents share a common ancestor, so it is possible that the heterozygous carriers share some other mutation causing short stature, and the

level of evidence of a heterozygous effect was limited by the size of the family. Based on these two studies, it is presumed that heterozygous NPR2 mutations can mildly impair long bone growth and it has further been hypothesized that one person in 30 with idiopathic short stature (ISS) will be carrier of an NPR2 mutation[6,12].

Recent studies searched for heterozygous NPR2 mutations in cohorts with ISS. One study of 47 independent Brazilian patients identified heterozygous NPR2 mutations in 6% of patients[13]. Another study on 101 unrelated Japanese patients with short stature identified heterozygous NPR2 mutations in 2% of patients[14]. While providing observational data consistent with the hypothesis that a monoallelic NPR2 mutation could cause short stature, these studies did not study normal height controls and were based on a relatively small number of patients, so have not rigorously verified the hypothesis. Analyses of larger cohorts, including controls, are needed to more clearly define the role of heterozygous NPR2 defects in patients with ISS.

RESULTS

Discovery of NPR2 variants in patients with short stature and controls using pooled sequencing

We selected 192 patients with short stature (>2 SD below the mean for age and sex) and 192 controls of matching ancestry from the middle of the Framingham Heart Study (FHS) height distribution (height z scores between -0.7 and +0.7 SDs). Characteristics of the subjects were described previously[15]. Pooled targeted sequencing of the exons of 1077 candidate genes was performed[15]. We detected 11 variants in NPR2. Of these, two were synonymous SNPs found in multiple patients and controls, and two were synonymous variants found only in patients. We focused on the seven potential loss-of-function variants, which included one splice site and 6

missense variants, all found in patients only (Table 3.1). These seven variants were all validated via traditional Sanger sequencing and confirmed to be heterozygous in each individual carrier.

Family analyses and clinical phenotypes

The splice site mutation is a C.1352-1G>A in a highly conserved base pair in the acceptor splice site at the 5' end of exon 7, carried in a single patient. Sanger sequencing of the proband, his parents, and a brother confirmed that the variant is found in the patient but not in his mother, father or brother. These results were confirmed by a second round of sequencing, and paternity was confirmed by SNP genotyping. There was no family history of short stature, consistent with the hypothesis that the *de novo* variant in NPR2 is contributing to short stature in this patient.

We also detected 6 missense variants in NPR2; of these, 4 were private variants not found in any of the reference databases (Patients 2-5, Table 3.1), and sequencing of relatives demonstrated segregation of these variants with the short stature phenotype (Figure 3.1). Of note, one of these variants (Patient 2) was found in a male patient who was also found to carry a known *TRPV4* mutation (c.1858G>A, V620I) previously reported as causal for brachyolmia type 3[16]. The patient carries a clinical diagnosis of brachyolmia and features consistent with this disease, including platyspondyly of the cervical spine, and we previously reported the presence of this likely pathogenic variant in this patient[15]. The NPR2 variant was present in the patient's mother, who does not carry the *TRPV4* variant, and in two sisters, the elder of which also carries the *TRPV4* variant. The mother and both sisters had short stature with height SDS scores below -2.5, and the sister carrying both variants had a height SDS score of -3.1. Notably patient's father is deceased but also was reported to have had short stature (-3.1 SDS) and presumably carried the *TRPV4* variant. Thus, this patient and one of his sisters inherited the NPR2 variant from his

mother and the *TRPV4* variant from his father, presumably resulting in severe short stature with skeletal dysplasia.

We also detected two rare missense variants present in the NHLBI Exome Sequencing Project data base (Patients 6-7). Neither of these variants has been reported to be pathogenic, and both are reported to be rare (minor allele frequency 0.01%). The mother of Patient 6, who also had short stature, did not carry the NPR2 variant, but DNA was not available from the sibling or father for further segregation analysis. Patient 7 was adopted from China, so the family history of short stature was unknown and testing for segregation of the variant was not possible.

Table 3.1: NPR2 potentially nonsynonymous variants in short stature patients

Subject	Sex	Variant	PolyPhen2 Prediction	Nadir Height SDS	Current Height SDS	GH Therapy	GH Indication	Upper/ Lower
Patient 1	M	Splice C.1352-1G>A	N/A	-3.14	-2.15	Yes	ISS	1.08 > +2 SDS
Patient 2	M	A48S (brachylomia with known TRPV4 variant)	1.00	-4.15	-3.91	No	N/A	1.17 > +2 SDS
Patient 3	F	E389D	0.00	-3.84	-3.14	Yes	ISS	0.91 -1 SDS
Patient 4	F	I494S	0.997	-3.57	-1.79	Yes	ISS	0.96 0 SDS
Patient 5	M	A549T	0.998	-3.88	-2.70	Yes	ISS	1.02 +1-2 SDS
Patient 6	M	A164G	0.065	-2.53	-2.16	IGF-1 therapy	ISS/IGF-1 deficiency	1.11 > +2 SDS
Patient 7	F	R787W	0.998	-2.92	-2.41	No	ISS	1.06 > +2 SDS

Abbreviations: F, female; M, male; GH, growth hormone; IGF-1, insulin-like growth factor 1.

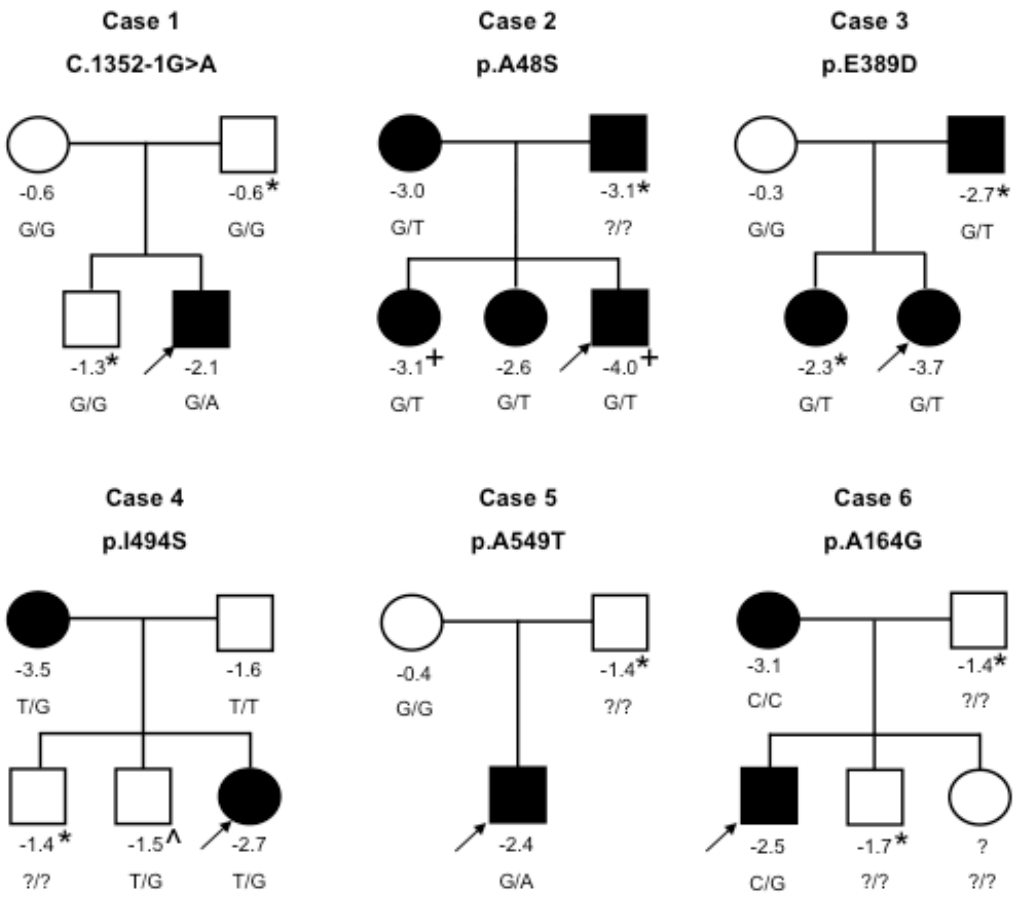

Figure 3.1: Segregation of identified NPR2 nonsynonymous variants in affected families. Numbers below the individuals denote the height z scores. + indicates family members who also carry the TRPV4 mutation casual of brachyolmia. ^ indicates that this family member was treated with growth hormone therapy prior to this height measurement and had a nadir height of -2.8 SDS. * indicates that the height was estimated by a family member. All other values were measured. Individuals with short stature are indicated as black circles (females) or squares (males). The arrow points to the affected proband in each family.

Screening for NPR2 mutations in cohorts of height extreme individuals

We then went on to screen for NPR2 mutations in two additional cohorts. The first cohort of individuals (n=272) was selected from the extremes (<1st percentile or >99th percentile) of height distribution from four FINRISK surveys (~33,000 samples in total). Pooled targeted sequencing of the exons of 1077 candidate genes was performed. We detected 8 potential variants in NPR2. Of these variants, two were common synonymous SNPs (observed in the short stature patients and FHS controls as well) and one was an intronic variant found in both short and tall extremes, and we did not pursue these further. Among the remaining 5 missense variants, 2 were validated by Sequenom genotyping and confirmed to be heterozygous in tall extremes only (Table 3.2). One of these is located in the extracellular region at the ligand binding domain (p.N247D) and the other is located at the kinase homology domain (p.R562Q). Neither variants were found in any of the reference databases. An in silico analysis suggested that one mutation is benign (p.N247D) and the other is probably damaging (p.R562Q).

The second cohort of individuals (n=1,000) was extremes (<1.25th percentile or >98.75th percentile) of height distribution from Estonian Biobank (~52,000 samples in total). Pooled targeted sequencing of the exons of four candidate genes was performed. Again, we did not evaluate further the synonymous or intronic variants identified. There were two missense variant observed in both tall extreme and short extreme samples, which we did not pursue further either. For the remaining 6 missense variants, we performed Sanger sequencing and validated 3 singleton variants (Table 3.2) all found in short extremes only. Two of them are located in the extracellular region at the ligand binding domain (p.E76K, p.G39R). The third mutation is located at the kinase homology domain (p.T661K). These allelic variants were not found in any

of the reference databases. An in silico analysis suggested that the three missense mutations are possibly damaging (p.E76K), benign (p.G39R), and probably damaging (p.T661K) respectively.

Table 3.2: NPR2 potentially nonsynonymous variants in FINRISK and Estonian GeneBank height extreme samples

Variant	Observation	PolyPhen2 Prediction
p.N247D	FINRISK tall extreme	0.002
p.R562Q	FINRISK tall extreme	0.995
p.E76K	Estonian short extreme	0.609
p.G39R	Estonian short extreme	0.013
p.T661K	Estonian short extreme	0.999

Overall observation of NPR2 mutations across three cohorts

Overall, across three cohorts, we observed 12 NPR2 nonsynonymous variants, 10 in short stature samples, and 2 in tall extreme samples (Table 3.3). Assuming that all variants are equally likely to occur in short stature samples and tall extreme samples (or control samples of normal height) under the null hypothesis, our observation gives a p value of 0.018 (one-tailed test).

We hypothesize that the 2 missense variants observed in FINRISK tall extremes are either functional neutral or gain-of-function mutations, while the missense variants identified in Estonian short extremes and in the seven short stature cases are loss-of-function mutations. If the hypothesis is validated, the association signal will be stronger than described above, as functional neutral variants can be removed from the test.

Table 3.3: Observation of NPR2 nonsynonymous variants in three cohorts

Cohort	**Observation of nonsynonymous variants**	**Note**
Short stature patients and FHS controls	7 variants in 192 short stature patients 0 variant in 192 FHS controls	Patient 1: de novo mutation Patients 2-6: all heterozygous relatives had short stature (height SDS < -2) Patient 7: no family data available
Extremes of FINRISK height distribution	0 variant in 136 short extremes 2 variants in 136 tall extremes	No family data available
Extremes of Estonian Biobank height distribution	3 variants in 500 short extremes 0 variant in 500 tall extremes	No family data available

Functional characterization of NPR2 variants

To further examine the pathogenicity of the identifed NPR2 nonysnonymous variants (Table 3.3), we initiated experiments assessing the CNP-dependent cGMP-producing capacities, and results are still pending. It will be interesting to see whether the results of functional studies are consistent with the outcome of familial analyses. We also intend to evaluate the correlation between in silico prediction of mutation effects with the results of functional studies.

Discussion

In summary, to explore the role of NPR2 variation in short stature, we have screened for heterozygous NPR2 mutations in three different cohorts (Table 3.3). Familial analyses in the short stature patient cohort support the hypothesis that rare heterozygous NPR2 variants could be a major contributor short stature in individuals carrying these variants. Functional studies of the NPR2 mutations identified are pending, which could add further support for this hypothesis. We also identified NPR2 mutations in two additional cohorts, where family data are not available. Pending results of functional studies will be crucial in interpreting these discoveries, especially as we identified two variants in individuals from the tall extreme of the height distribution. The frequencies of heterozygous mutation carriers in varied across cohorts. This is likely due to different ethnicities, sample selection criteria, and possibly experimental design.

In contrast with homozygous mutations in NPR2, which produces a severe short stature and body disproportion, heterozygous mutations in NPR2 seem to be associated with mild and variable growth impairment without distinct phenotype. The severity of short stature, body proportions, and the presence of nonspecific skeletal abnormalities vary across individuals in our

study, consistent with previous observations[12,13]. This variability is likely due to differences in the nature of NPR2 mutations carried by the individuals, as well as variable expressivity.

Previously, a study of 47 Brazilian patients identified heterozygous NPR2 mutations in 6% of patients[13], while another study on 101 short stature Japanese patients identified mutations in 2% of patients[14]. Functional analyses were performed in both studies to evaluate the pathogenicity and elucidate the molecular mechanisms of the identified mutations. However, both studies were observations based on small number of patients. Our study, in much larger cohorts and with more family data, with functional analyses (once completed), will be a more rigorous assessment of heterozygous mutations in NPR2 as a potential cause of growth impairment in ISS patients.

The frequency of heterozygous carriers of AMDM mutations was previously estimated to be ~0.14%[12]. In NHLBI, the cumulative frequency of all nonsynonymous NPR2 variants is approximately 0.4% in ~4000 European American samples. The discrepancy between these two allele frequencies highlights the importance of functional studies, which boost the power of rare variant association tests by removing neutral background variants. Despite the rarity of AMDM mutations (and nonsynonymous NPR2 variants in general), these loss-of-function alleles likely have relatively large effect size (previously estimated to be around -1.8 SDS[12]) and would explain ~1% of height variation in the population based on these assumptions. Moreover, common variants at the locus and in this pathway also contribute to height variation, as GWAS signals have been identified near both NPR2 and *NPPC*, encoding the ligand CNP.

The action of natriuretic peptide system on longitudinal growth is partially explained by the capacity of CNP/NPR2 signaling to inhibit fibroblast growth factor receptor-3 (FGFR3) downstream signaling at the level of the MAPK cascade[17]. Gain-of-function mutations in

FGFR3, which promote a sustained activation of the MAPK pathway, are responsible for achondroplasia, one of the most common skeletal dysplasia[18]. It has been shown that CNP alleviates the short-limbed phenotype of achondroplasia mice[19]. The CNP analog with an extended half-life, BMN 111, has recently been developed and significantly recovery of bone growth was demonstrated in ACH mice by subcutaneous administration of BMN 111[20].

While this therapeutic approach would be ineffective in AMDM patients that lack the receptor for CNP, this form of therapy might be effective for heterozygous carriers of NPR2 mutations who still have one functional receptor. A study by Olney et al. has shown that plasma CNP level was very high in AMDM patients, suggesting the presence of a feedback loop that regulates CNP production[12]. For the heterozygous carriers of the mutation, this level was not different from the noncarriers. In the normal population, blood level of CNP is lower in adults than in children. It is therefore also possible that NPR2 mutation carriers have abnormal level of CNP during childhood but normal level in adulthood. This finding would likely not have been detected in the study of Olney et al. because most of the participants were adults.

Apart from CNP, it is conceivable that pharmacological inhibitors of the MEK/ERK MAPK pathway might improve bone growth for AMDM patients and heterozygous carriers. MEK1/2 inhibitors that are currently undergoing clinical evaluation in cancer are promising therapeutic candidates for growth defects associated with excess MEK1/2 signaling, including patients with AMDM, achondroplasia or other appropriate skeletal dysplasia. It has been shown that pharmacological inhibition of MEK1/2 is sufficient to rescue the growth defect in mice model of AMDM[21]. The key to moving skeletal dysplasia therapeutics forward will be to deliver this class of agents specifically to the growth plate.

AMDM patients appear to have an abnormality in the GH/IGF-1 system, characterized by low insulin-like growth factor 1 (IGF-1) levels, high growth hormone (GH) levels, and lack of a response to GH treatment[12]. This implies an interaction between CNP/NPR2 and GH/IGF-1 pathways during postnatal growth and the growth failure in AMDM may be due at least partially to low IGF-1 levels. Studies in more children with AMDM will test this hypothesis. IGF-1 levels, however, were not low in the carriers[12–14]. It is possible, as mentioned earlier, that carriers have abnormal levels of IGF-1 during childhood that were not detected. Studies that will advance our knowledge of these mechanisms can eventually help predict the determinants for responsiveness to therapy. This knowledge would allow clinicians to tailor GH treatment and dosing to an individual's molecular diagnosis. It may be useful in the future to include NPR2 sequencing in the evaluation of children presenting with short stature. This information could then be used to help make the decision of whether to start growth hormone.

MATERIALS AND METHODS

Subjects

Short stature patient cohort

192 patients with short stature (>2 SD below the mean for age and sex[22]) but without defined genetic etiologies, were recruited from the endocrinology and genetics clinics at Boston Children's Hospital. In addition, 192 control subjects were chosen from the middle of the Framingham Heart Study height distribution (height z scores between −0.7 and +0.7 SDs). More detailed description of the cohort is reported in Wang et al.[15]

FINRISK height extreme cohort

272 subjects were chosen from the extremes (<1st percentile or >99th percentile) of the FINRISK surveys (FINRISK 1992, FINRISK 1997, FINRISK 2002, FINRISK 2007) height distribution. The FINRISK cohorts comprise the respondents of representative, cross-sectional population surveys that are carried out every 5 years since 1972, to assess the risk factors of chronic diseases and health behavior in the working age population, in five large study areas of Finland[23].

Estonian Biobank height extreme cohort

1000 subjects were selected from the extremes (<1.25th percentile or >98.75th percentile) of the Estonian Biobank height distribution. The Estonian Biobank cohort is a volunteer-based sample of the Estonian resident adult population (age ≥18 years). The age, sex and geographical distribution closely reflect those of the Estonian adult population and encompass close to 5% of the entire adult population of Estonia[24].

Pooled sequencing protocol

DNA samples from multiple subjects were pooled for DNA sequencing using previously described methods available at the Broad Institute[25]. To identify variants present only in a single individual (hereon referred to as singleton variants), we applied a simple overlapping pooling design as described in Wang et al.[15]. In short, each sample was sequenced in 2 pools (1 row pool and 1 column pool). Singleton variants appear only in 1 row pool and 1 column pool. Therefore, the subject whose DNA sample is present at the intersection of these 2 pools must be the individual carrying that singleton variant.

Sequencing was performed on the Illumina HiSeq platform. Variant calling was performed as described previously[15]. Variants were annotated for functional effect using SnpEff 2.0.5 (http://snpeff.sourceforge.net/). Variant allele frequency data were obtained from the National Heart, Lung, and Blood Institute Exome Variant Server[26].

Confirmation of variants found in NPR2 through pooled sequencing was done via Sanger sequencing or Sequenom genotyping. Each variant was sequenced in the proband and in all related family members for whom DNA samples were provided.

Short stature patient cohort

The samples from short stature subjects and control samples were each arranged into a 14 × 14 matrix of 28 pools, with 13 to 14 samples in each pool. The coding regions of NPR2 were sequenced along with ~1000 height candidate genes[15]. The targeted exons of the candidate genes were enriched using a custom Agilent SureSelect hybrid selection system. The mean target coverage of NPR2 is 297 reads per pool, resulting in 21 reads per subject in a pool of 14 subjects or 42 total reads per subject, as each subject is present in 2 pools.

FINRISK height extreme cohort

The FINRISK DNA samples were whole-genome amplified. The short extreme subjects and the tall extreme subjects were each arranged into a 12 × 12 matrix of 24 pools, with 11-12 samples in each pool. The coding region of NPR2 was enriched as described above for short stature patient cohort. The mean target coverage of NPR2 is 377 reads per pool, resulting in 31 reads per subject in a pool of 12 subjects or 62 total reads per subject.

Estonian Biobank height extreme cohort

The samples from short extreme subjects and tall extreme subjects were each arranged into a 24 × 24 matrix of 48 pools, with 19-24 samples in each pool. The coding regions of NPR2 were sequenced along with three other genes. The targeted exons of all four genes were enriched using PCR-based fluidigm Access Array system. The mean target coverage of NPR2 is 2735 reads per pool, resulting in 114 reads per subject in a pool of 24 subjects or 228 total reads per subject.

In silico prediction of mutation effects

To identify the potential effects of sequence variants identified in NPR2 on protein function or structure, the wild-type and variant sequences were submitted to the PolyPhen method (http://genetics.bwh.harvard.edu/pph2)[27].

Assaying wild-type and mutant NPR2 activity

Missense mutations in NPR2 will be generated by site-directed mutagenesis using the wild-type rat NPR2 expression construct pRK-NPR-B. Activity in HEK 293 cells will be measured as described elsewhere[28]. In brief, cells will be seeded to 40%-45% confluency in 10-cm dishes in Dulbecco's modified Eagle medium (DMEM) with 10% fetal bovine serum (FBS). The cells will be transfected 2 d later with 4 μg of expression construct with Lipofectamine (Invitrogen) in serum-free DMEM. After transfection, cells will be supplemented with 15% FBS and incubated overnight. Transfection efficiency is expected to be 40-50%, as determined by cotransfecting cells with a green fluorescent protein reporter plasmid. Transfected cells will be starved for 12 h prior to CNP exposure (48-72 h after transfection). Cells will then be exposed to

1μM CNP (Bachem) for 3 min. Signaling will be terminated by aspirating the CNP-containing media and adding ice-cold 80% ethanol to the cells. This ethanol extract will be centrifudged, and the supernatant will be collected and vacuum evaporated. The amount of cGMP in the evaporated samples will be measured by use of a cGMP assay kit (Cayman Chemical).

References

1. Schulz, S. (2005). C-type natriuretic peptide and guanylyl cyclase B receptor. Peptides *26*, 1024–1034.

2. Potter, L.R., Abbey-Hosch, S., and Dickey, D.M. (2006). Natriuretic peptides, their receptors, and cyclic guanosine monophosphate-dependent signaling functions. Endocr. Rev. *27*, 47–72.

3. Yasoda, A., Komatsu, Y., Chusho, H., Miyazawa, T., Ozasa, A., Miura, M., Kurihara, T., Rogi, T., Tanaka, S., Suda, M., et al. (2004). Overexpression of CNP in chondrocytes rescues achondroplasia through a MAPK-dependent pathway. Nat. Med. *10*, 80–86.

4. Chusho, H., Tamura, N., Ogawa, Y., Yasoda, A., Suda, M., Miyazawa, T., Nakamura, K., Nakao, K., Kurihara, T., Komatsu, Y., et al. (2001). Dwarfism and early death in mice lacking C-type natriuretic peptide. Proc. Natl. Acad. Sci. U. S. A. *98*, 4016–4021.

5. Tamura, N., Doolittle, L.K., Hammer, R.E., Shelton, J.M., Richardson, J.A., and Garbers, D.L. (2004). Critical roles of the guanylyl cyclase B receptor in endochondral ossification and development of female reproductive organs. Proc. Natl. Acad. Sci. U. S. A. *101*, 17300–17305.

6. Bartels, C.F., Bükülmez, H., Padayatti, P., Rhee, D.K., van Ravenswaaij-Arts, C., Pauli, R.M., Mundlos, S., Chitayat, D., Shih, L.-Y., Al-Gazali, L.I., et al. (2004). Mutations in the transmembrane natriuretic peptide receptor NPR-B impair skeletal growth and cause acromesomelic dysplasia, type Maroteaux. Am. J. Hum. Genet. *75*, 27–34.

7. Moncla, A., Missirian, C., Cacciagli, P., Balzamo, E., Legeai-Mallet, L., Jouve, J.-L., Chabrol, B., Le Merrer, M., Plessis, G., Villard, L., et al. (2007). A cluster of translocation breakpoints in 2q37 is associated with overexpression of NPPC in patients with a similar overgrowth phenotype. Hum. Mutat. *28*, 1183–1188.

8. Bocciardi, R., Giorda, R., Buttgereit, J., Gimelli, S., Divizia, M.T., Beri, S., Garofalo, S., Tavella, S., Lerone, M., Zuffardi, O., et al. (2007). Overexpression of the C-type natriuretic peptide (CNP) is associated with overgrowth and bone anomalies in an individual with balanced t(2;7) translocation. Hum. Mutat. *28*, 724–731.

9. Miura, K., Kim, O.-H., Lee, H.R., Namba, N., Michigami, T., Yoo, W.J., Choi, I.H., Ozono, K., and Cho, T.-J. (2014). Overgrowth syndrome associated with a gain-of-function mutation of the natriuretic peptide receptor 2 (NPR2) gene. Am. J. Med. Genet. A *164A*, 156–163.

10. Robinson, J.W., Dickey, D.M., Miura, K., Michigami, T., Ozono, K., and Potter, L.R. (2013). A human skeletal overgrowth mutation increases maximal velocity and blocks desensitization of guanylyl cyclase-B. Bone *56*, 375–382.

11. Miura, K., Namba, N., Fujiwara, M., Ohata, Y., Ishida, H., Kitaoka, T., Kubota, T., Hirai, H., Higuchi, C., Tsumaki, N., et al. (2012). An overgrowth disorder associated with excessive production of cGMP due to a gain-of-function mutation of the natriuretic peptide receptor 2 gene. PLoS One *7*, e42180.

12. Olney, R.C., Bükülmez, H., Bartels, C.F., Prickett, T.C.R., Espiner, E.A., Potter, L.R., and Warman, M.L. (2006). Heterozygous mutations in natriuretic peptide receptor-B (NPR2) are associated with short stature. J. Clin. Endocrinol. Metab. *91*, 1229–1232.

13. Vasques, G.A., Amano, N., Docko, A.J., Funari, M.F.A., Quedas, E.P.S., Nishi, M.Y., Arnhold, I.J.P., Hasegawa, T., and Jorge, A.A.L. (2013). Heterozygous Mutations in Natriuretic Peptide Receptor-B (NPR2) Gene as a Cause of Short Stature in Patients Initially Classified as Idiopathic Short Stature. J. Clin. Endocrinol. Metab. *98*, E1636–E1644.

14. Amano, N., Mukai, T., Ito, Y., Narumi, S., Tanaka, T., Yokoya, S., Ogata, T., and Hasegawa, T. (2014). Identification and Functional Characterization of Two Novel NPR2 Mutations in Japanese Patients with Short Stature. J. Clin. Endocrinol. Metab. jc20133525.

15. Wang, S.R., Carmichael, H., Andrew, S.F., Miller, T.C., Moon, J.E., Derr, M.A., Hwa, V., Hirschhorn, J.N., and Dauber, A. (2013). Large Scale Pooled Next-Generation Sequencing of 1077 Genes To Identify Genetic Causes of Short Stature. J. Clin. Endocrinol. Metab.

16. Rock, M.J., Prenen, J., Funari, V.A., Funari, T.L., Merriman, B., Nelson, S.F., Lachman, R.S., Wilcox, W.R., Reyno, S., Quadrelli, R., et al. (2008). Gain-of-function mutations in TRPV4 cause autosomal dominant brachyolmia. Nat. Genet. *40*, 999–1003.

17. Krejci, P., Masri, B., Fontaine, V., Mekikian, P.B., Weis, M., Prats, H., and Wilcox, W.R. (2005). Interaction of fibroblast growth factor and C-natriuretic peptide signaling in regulation of chondrocyte proliferation and extracellular matrix homeostasis. J. Cell Sci. *118*, 5089–5100.

18. Vajo, Z., Francomano, C.A., and Wilkin, D.J. (2000). The molecular and genetic basis of fibroblast growth factor receptor 3 disorders: the achondroplasia family of skeletal dysplasias, Muenke craniosynostosis, and Crouzon syndrome with acanthosis nigricans. Endocr. Rev. *21*, 23–39.

19. Yasoda, A., Kitamura, H., Fujii, T., Kondo, E., Murao, N., Miura, M., Kanamoto, N., Komatsu, Y., Arai, H., and Nakao, K. (2009). Systemic administration of C-type natriuretic peptide as a novel therapeutic strategy for skeletal dysplasias. Endocrinology *150*, 3138–3144.

20. Lorget, F., Kaci, N., Peng, J., Benoist-Lasselin, C., Mugniery, E., Oppeneer, T., Wendt, D.J., Bell, S.M., Bullens, S., Bunting, S., et al. (2012). Evaluation of the therapeutic potential of a CNP analog in a Fgfr3 mouse model recapitulating achondroplasia. Am. J. Hum. Genet. *91*, 1108–1114.

21. Geister, K.A., Brinkmeier, M.L., Hsieh, M., Faust, S.M., Karolyi, I.J., Perosky, J.E., Kozloff, K.M., Conti, M., and Camper, S.A. (2013). A novel loss-of-function mutation in Npr2 clarifies primary role in female reproduction and reveals a potential therapy for acromesomelic dysplasia, Maroteaux type. Hum. Mol. Genet. *22*, 345–357.

22. Kuczmarski, R.J., Ogden, C.L., Guo, S.S., Grummer-Strawn, L.M., Flegal, K.M., Mei, Z., Wei, R., Curtin, L.R., Roche, A.F., and Johnson, C.L. (2002). 2000 CDC Growth Charts for the United States: methods and development. Vital Health Stat. 11. 1–190.

23. Vartiainen, E., Laatikainen, T., Peltonen, M., Juolevi, A., Männistö, S., Sundvall, J., Jousilahti, P., Salomaa, V., Valsta, L., and Puska, P. (2010). Thirty-five-year trends in cardiovascular risk factors in Finland. Int. J. Epidemiol. *39*, 504–518.

24. Leitsalu, L., Haller, T., Esko, T., Tammesoo, M.-L., Alavere, H., Snieder, H., Perola, M., Ng, P.C., Mägi, R., Milani, L., et al. (2014). Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. Int. J. Epidemiol.

25. Rivas, M.A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C.K., Boucher, G., Ripke, S., Ellinghaus, D., Burtt, N., et al. (2011). Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat. Genet. *43*, 1066–1073.

26. National Heart, Lung, and B.I. Exome Variant Server.

27. Ramensky, V., Bork, P., and Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. Nucleic Acids Res. *30*, 3894–3900.

28. Abbey, S.E., and Potter, L.R. (2003). Lysophosphatidic acid inhibits C-type natriuretic peptide activation of guanylyl cyclase-B. Endocrinology *144*, 240–246.

Chapter 4

Simulation of Finnish population history, guided by empirical genetic data, to assess power of rare variant tests in Finland

Sophie R. Wang[1,2,3], Vineeta Agarwala[2,4,5], Jason Flannick[2,6], Charleston W.K. Chiang[7], David Altshuler[2,3,6,8], GoT2D Consortium, Joel N. Hirschhorn[1,2,3,]

1. Division of Endocrinology, Boston Children's Hospital, Boston, MA 02115, USA
2. Program in Medical and Population Genetics, Broad Institute, Cambridge, MA 02142, USA
3. Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
4. Harvard-MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
5. Program in Biophysics, Graduate School of Arts and Sciences, Harvard University, Cambridge, MA 02138, USA
6. Department of Molecular Biology, Massachusetts General Hospital, Boston, MA 02114, USA
7. Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095, USA
8. Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Originally published as:
Wang SR, et al. *Simulation of Finnish population history, guided by empirical genetic data, to assess power of rare-variant tests in Finland*. Am J Hum Genet, 2014. **94**(5): p.710-20.

ABSTRACT

Finnish samples have been extensively utilized in studying single-gene disorders, where the founder effect has clearly aided in discovery, and more recently in genome-wide association studies (GWAS) of complex traits, where the founder effect has had less obvious impacts. As the field starts to explore rare variants' contribution to polygenic traits, it is of great importance to characterize and confirm the Finnish founder effect in sequencing data and to assess its implications for rare variants association studies. Here, we employ forward simulation, guided by empirical deep resequencing data, to model the genetic architecture of quantitative polygenic traits in both the general European and the Finnish population simultaneously. We demonstrate that power for rare variant association tests is higher in the Finnish population, especially when variants' phenotypic effects are tightly coupled with fitness effects and therefore reflect a greater contribution of rarer variants. SKAT-O, VT, and single variant tests are more powerful than other rare variant methods in the Finnish population across a range of genetic models. We also compare the relative power and efficiency of exome array genotyping *vs.* high coverage exome sequencing. At a fixed cost, less expensive genotyping strategies have far greater power than sequencing; in a fixed number of samples, however, genotyping arrays miss a substantial portion of genetic signals detected in sequencing, even in the Finnish founder population. As genetic studies probe sequence variation at greater depth in more diverse populations, our simulation approach provides a framework to evaluate various study designs for gene discovery.

INTRODUCTION

A founder effect can result either from a true founder event (i.e., the establishment of a new population from a limited pool of individuals) or from an extreme reduction in population size (i.e., a bottleneck in size), followed by relative genetic isolation from other populations. The population of Finland is one of the best-studied genetic isolates. The Finnish genetic architecture has been shaped by a series of founder effects and a subsequent drift in local subisolates. The initial founder effects are generally associated with two colonization waves 4000 and 2000 years ago to southern and western Finland. More recently, there was an internal migration movement in the 15th-16th century from a small southeastern area to the middle, western and finally northern and eastern parts of the country[1].

The Finnish population has been extensively utilized in genetic studies. It is considered to be a relatively homogenous large founder population, and hence potentially well suited for genetic mapping. The evidence for a founder effect includes enrichment of almost forty rare recessive diseases, longer regions of linkage disequilibrium, increased kinship coefficients between pairs of randomly chosen individuals, and extended runs of homozygosity[1–10]. In part because of the founder effect, identification of the genes underlying the rare diseases enriched in Finland has been remarkably successful.[1] Finnish samples have also contributed to many GWAS of complex traits, but because common (>5% minor allele frequency, MAF) variation is less influenced by human population history, a founder effect would be less likely to provide a specific advantage in this setting.

Studies of polygenic traits and disorders are now moving to a middle ground between GWAS genotyping methods (thus far focused largely on common variation, typically MAF >5%) and sequencing-based methods that were most successfully employed to identify extremely rare

variants in single gene disorders. This middle ground is association studies of lower frequency variants (MAF <5%), analyzed either individually or in aggregate (the aggregate analysis has also been termed RVAS[11]). As such, it is of great interest and importance to confirm the Finnish founder effect in sequencing data that includes rarer variants and assess the implications for these association studies of rarer variants. Due to the founding event and subsequent strong genetic drift, some variants that are rare in the ancestral population will have risen in frequency in a founder population, while others will have decreased or disappeared. These alterations in allele frequency could potentially increase power of rare variant tests in two ways. First, some rare and potentially deleterious variants could rise to higher frequencies, out of proportion to what might be expected given their deleterious effects. Second, there is greater homogeneity of rare variation in a founder population, and thus fewer background rare variants at any individual locus. As an example, a protective mutation for Alzheimer's disease (MIM 104300) was discovered in part because it has a much higher frequency in the Scandinavian populations (~0.4%) than in the general European population (<0.01%)[12].

Exome sequencing studies are emerging as a popular approach to identify rare coding variants associated with complex traits, while a cheaper alternative approach is to use array-based genotyping of a defined set of coding variants. The human genetics community has aggregated an extensive list of putative functional coding variants from the exome sequences of >12,000 individuals for array-based genotyping platforms (e.g. the Illumina Infinium HumanExome BeadChip and the Affymetrix Axiom Exome Array Plate; see URL for a description of SNP content and selection strategies). Although these arrays do not provide a complete catalogue of all coding variants, the set of variants selected for array design is estimated to include >97% of the non-synonymous variants that would be detected in any

individual genome through exome sequencing. In theory, the coverage would be even higher for a founder population, which has fewer rare variants compared to a non-founder population.

To increase power to detect effects of rare variants, especially those that are too infrequent to be individually tested for association, many groups have devised tests that combine evidence across multiple variants[13]. These tests have become a standard approach for analyzing rare variants, and include burden tests[14–16] and other types of tests that aggregate evidence across sets of variants[17,18]. The relative power of such tests to detect association is strongly influenced by underlying genetic architecture. Specifically, the proportion of causal variants among all variants analyzed and the distribution of effect sizes and allele frequencies of causal variants all affect test performance. Different statistical tests also have different sets of parameters, the values of which can have a large effect on the power of the tests.

Different diseases and phenotypes likely have different architectures[19]. To try to evaluate how different sample selections (founder *vs.* non-founder populations), analytical methods (single variant tests, gene-based tests), and study designs (exome sequencing *vs.* exome array genotyping) would perform with different genetic architectures, we have developed a population genetics framework to assess the impact of the Finnish population history on genetic studies of rarer variation. Our approach has four basic stages: 1) confirming and characterizing the Finnish founder effect in sequence data, 2) developing a simultaneous simulation of sequence variation in the non-Finnish European (NFE) population and the Finnish population that closely approximates the sequence data, 3) specifying a range of models of genetic architectures to generate simulated phenotypic data, and 4) comparing operating characteristics of different gene-based tests and single variant tests on phenotype, genotype, and sequence data from simulated founder and non-founder populations.

With this framework in place, we address the following questions: 1) Under what types of genetic architecture(s) is it more powerful to use a founder population such as Finland? 2) Under different genetic models, what are the optimal association tests for rare variants in a founder population? 3) How does power compare between using exome sequencing data and exome chip data, particularly in a founder population? Our results show that power to detect genetic signals – by both single variant and gene-based tests – is higher in samples from the Finnish founder population than in equivalently sized NFE samples, especially when the phenotypic effects of variants are tightly coupled with effects on fitness. SKAT-O, VT, and single variant tests have the highest mean power in a founder population across simulated datasets. At a fixed cost, genotyping strategies have far greater power than sequencing; in a fixed number of samples, however, genotyping arrays miss a substantial portion of causal variation detected in sequencing.

RESULTS

Assessing the Finnish founder effect

We analyzed whole exome sequence data of NFE and Finnish samples from the GoT2D Project (see Methods). If the Finnish population had been through a founding event in the past, there are a number of direct predictions for the allele frequency spectra (AFS) and the sharing of variants between the Finns and the NFEs. We showed that, when comparing Finnish and NFE samples of the same size (N=500), the allele frequency spectra (AFS) are shifted towards higher frequencies in the Finns (Figure 4.1). The proportion of singleton variants is much lower in the Finns than the NFEs (28% *vs.* 46% for synonymous variants, 39% *vs.* 57% for missense variants), while the opposite is true for common variants (>5%) (31% *vs.* 22% for synonymous variants, 19% *vs.* 12% for missense variants). Furthermore, singleton variants in a population of

Finns (N=250) are more likely to be seen again in another population of Finns (N=250), compared with the same analysis in NFEs (34% *vs.* 23% for synonymous singleton variants, 30% *vs.* 20% for missense singleton variants) (Figure S4.1). We also observed that SNPs found in both samples tend to have higher frequencies in Finns (paired t-test p value < 0.01) (Table S4.1). Compared to the NFEs, the Finns have lower level of heterozygosity (on average there are 0.6% fewer heterozygous sites per individual in the Finns, t-test p value < 0.01) and reduced genetic diversity (Watterson's estimate adjusted by sequence length is 6.14×10^{-4} for the Finns and 1.01×10^{-3} for the NFEs). All these results strongly confirm the presence of a founder effect in Finland.

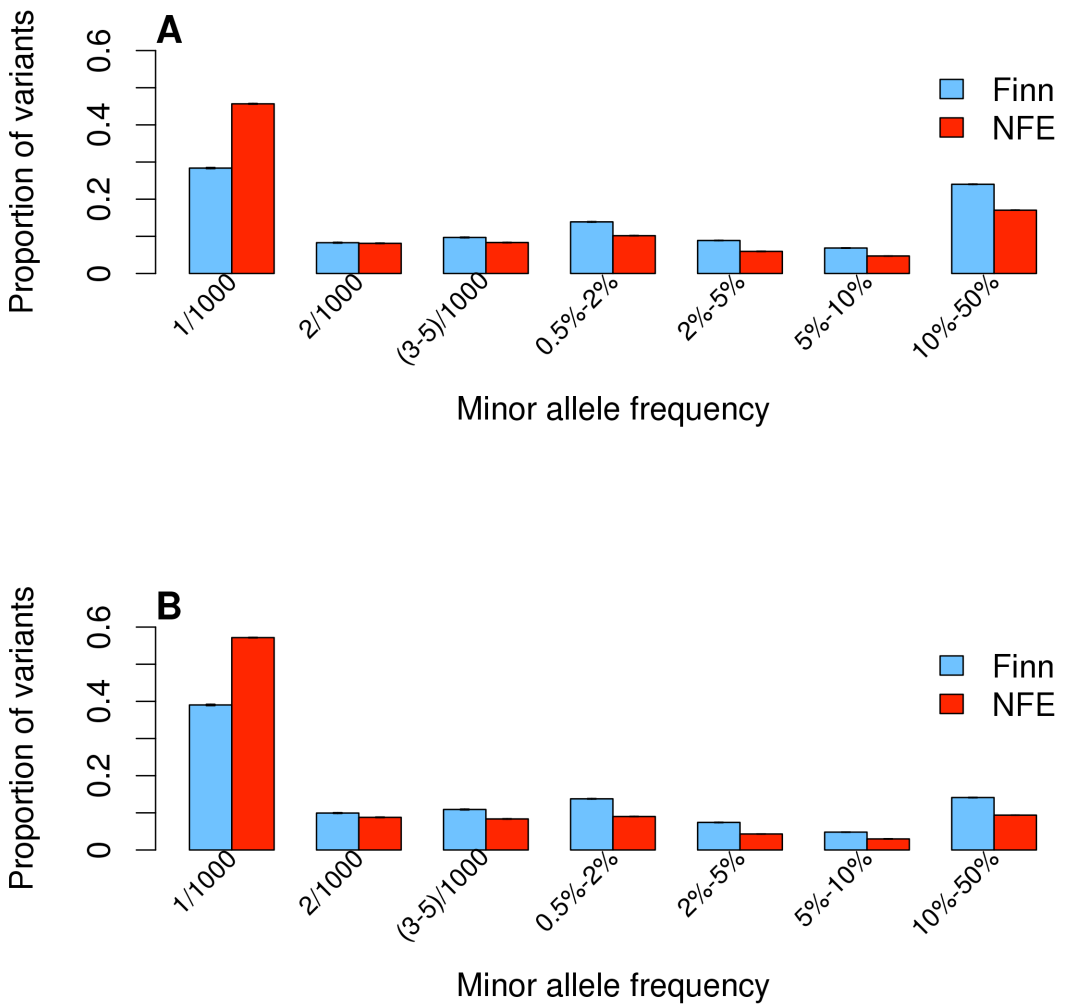

Figure 4.1: The final demographic model for simulating NFEs and Finns simultaneously. The NFEs were modeled as long-term (45,000 generations) constant size (N=8,100) followed by a bottleneck (N=2,000) and then by exponential growth (1.5% growth per generation). For modeling the Finns, we tested three general classes of models, of which only one (Class 3 in Tables S2) approximated the empirical observations. In this model, after the initial founding event (100 generations ago, N=1,000), the Finn-s went through a slow growth phase (0.5-5% growth per generation), and then a more recent fast growth phase (8-30% growth per generation); there was gene flow from the NFEs to the Finns.

Simulation of coding sequence variation in hundreds of thousands of samples

To enable a controlled characterization of the performance of different sample selections, analytical methods and study designs under a range of scenarios, we used the forward simulation package ForSim[20] to generate coding sequence data for the NFEs and the Finns simultaneously. This way, we could simulate evolution of complex traits over time in large samples and we know the truth (i.e. fitness effects) about all variants. To model the NFEs, we used the conventional four-parameter model[21] and adapted parameters from a recent simulation that generated representative sequence data for European populations (Figure 4.2)[22]. We further modeled the Finns as a founder population established by a small number of NFEs. We refined our demographic parameters for the Finnish model by comparing to exome sequencing data from the GoT2D project (see Method). In our final model, the initial founding event was followed by a slow growth phase, and then a more recent fast growth phase, with gene flow from the NFEs to Finns (Figure 4.2). Figure 4.3 shows that our final demographic model reproduces the observed allele frequency spectra well. We also analyzed missense/synonymous ratio (Figure S4.2) and allele sharing between the Finns and the NFEs (Figure S4.3); these metrics are also similar between the observed and simulated data.

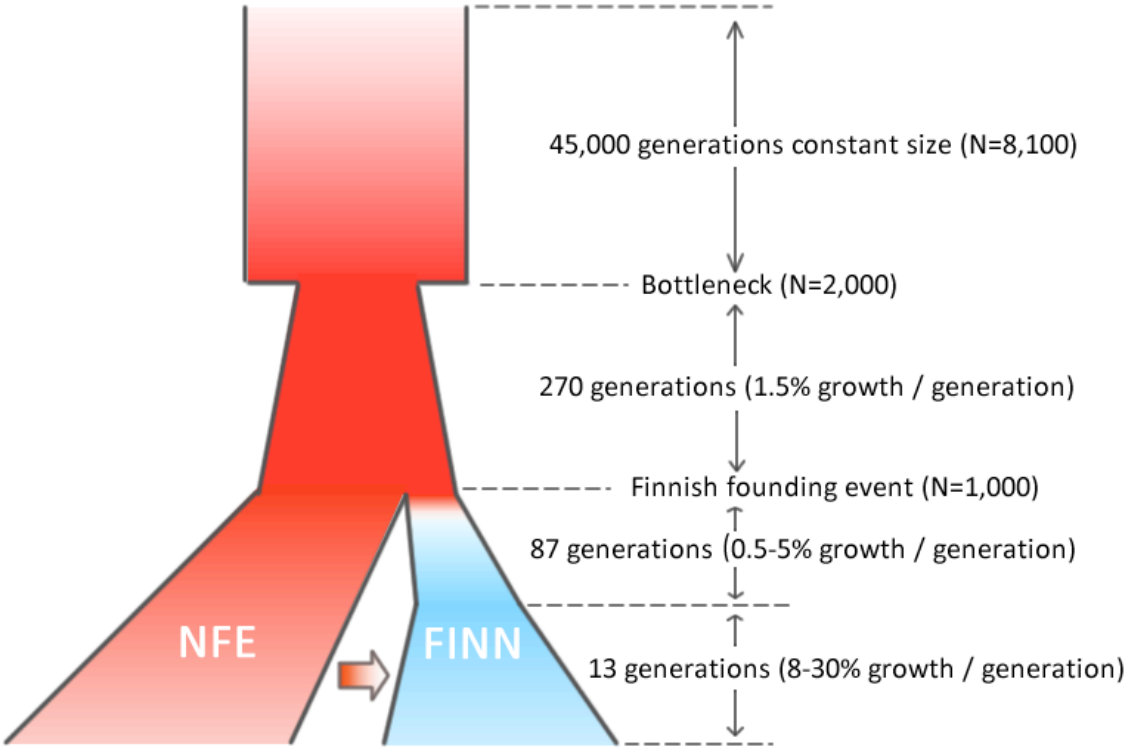

Figure 4.2: The final demographic model for simulating NFEs and Finns simultaneously. The NFEs were modeled as long-term (45,000 generations) constant size (N=8,100) followed by a bottleneck (N=2,000) and then by exponential growth (1.5% growth per generation). For modeling the Finns, we tested three general classes of models, of which only one (Class 3 in Tables S2) approximated the empirical observations. In this model, after the initial founding event (100 generations ago, N=1,000), the Finns went through a slow growth phase (0.5-5% growth per generation), and then a more recent fast growth phase (8-30% growth per generation); there was gene flow from the NFEs to the Finns.

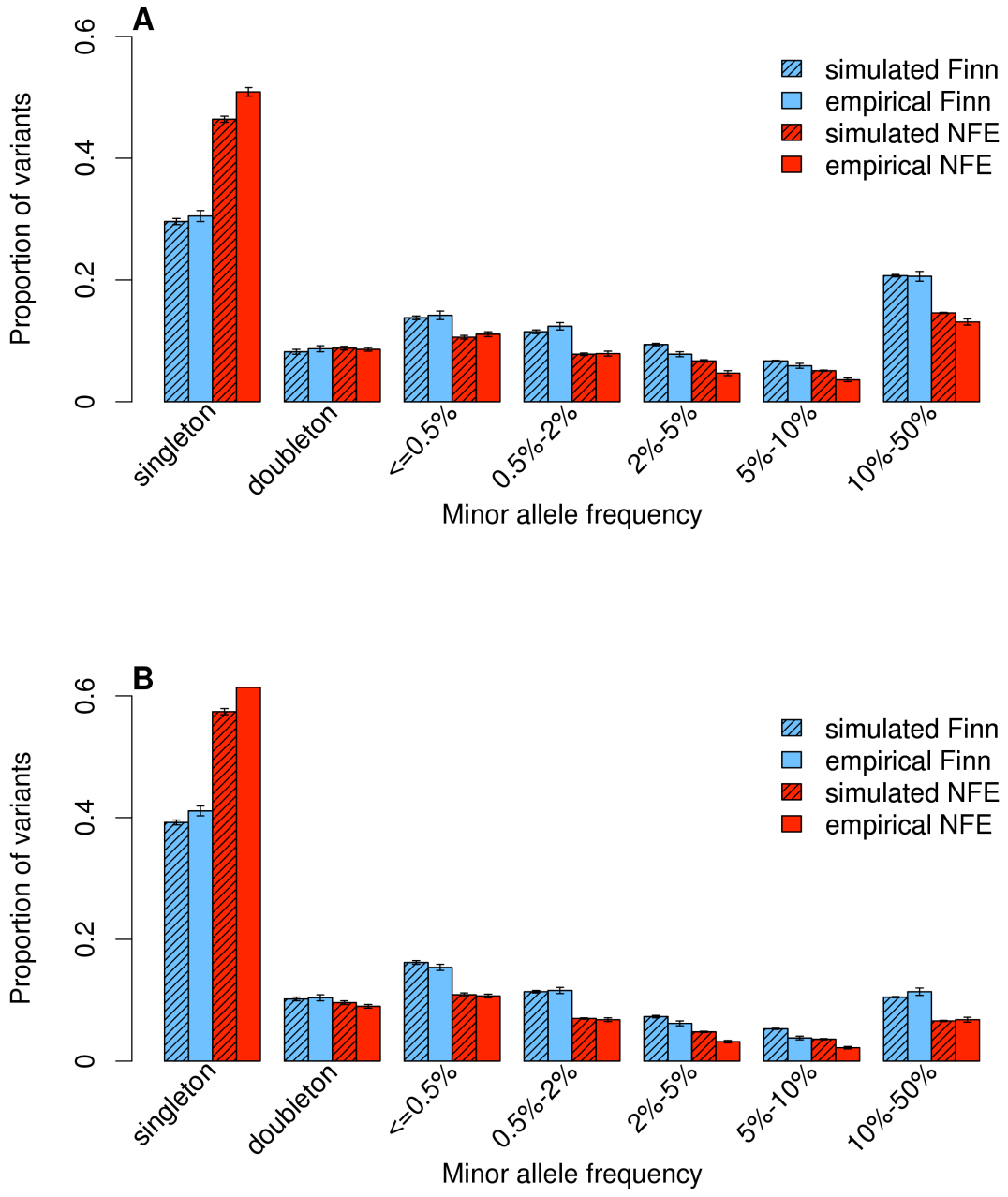

Figure 4.3: Agreement of empirical allele frequency spectra with the modeled spectra. Sample sizes are 843 for the Finns and 820 for the NFEs. (A) Synonymous variants and (B) Missense variants.

Specification of a range of disease models

Protein-coding variation will only partially explain the phenotypic variation of any polygenic trait. However, to focus on the role of coding variation (as they are more likely to enrich for functionally significant alleles), we simulated a heritable quantitative trait (h^2 = 80%) for which aggregated coding variation in each of 1,000 genes explains, on average, 0.1% of total heritability. We assume selectively neutral missense variants are background variants with no effects on the trait, while selectively non-neutral missense variants are the causal variants. Four different disease models were generated by varying the degree of coupling (τ) between a causal variant's phenotypic effect and the strength of purifying selection against that variant[23]. Broadly, M1 (τ=0) is characterized by rare and common alleles that have similar effects on phenotype; M2 (τ=0.5) produces a modest correlation between variant frequency and effect size; and M3 (τ=1) results in a sharp inverse correlation. M4 (τ is randomly chosen among 0, 0.5 and 1 for each effect gene) may represent a more realistic scenario, as different genes are likely to have different pleiotropic effects and are therefore exposed to different strengths of purifying selection. As expected, we observed that as τ increases, more phenotypic variance is explained by rare variants (Figure S4.4).

Alterations in allele frequency in the founder population

With simulated data we demonstrated alterations in allele frequency in the founder population, which could potentially increase power of rare variant tests. We first showed that there is greater homogeneity of rare variation at any individual locus in a founder population. The Finns have on average 2.5x fewer rare variants (<5%) per gene compared to the NFEs (mean 20.0 ± 4.5 *vs.* 52.3 ± 7.4). This reduction in rare variants was seen for both variants we

simulated as causal and those simulated as neutral, background variants (Figure S4.5). As seen in Figure S4.6, the cumulative allele frequency of causal variants and background variants per gene is similar between the Finns and the NFEs, meaning that there are fewer rare variants in the Finns, but they are each on average more common than variants in NFEs.

We next showed that there is increased frequency of causal variants (thus variance explained per gene) at some genes. We observed that the distribution of the variance explained per gene is wider in the Finns than in the NFEs (Figure 4.4). At one end of the distribution (the left tails of the graphs in Figure 4.4), the increased frequency of some individual causal variants leads to a greater variance explained for some genes in the Finns (more obvious with larger τ); at the other end of the distributions, so many causal variants are lost in Finland so other genes have lower variance explained in the Finns. As a result, some genes will be detectable in smaller sample sizes in Finns than in NFEs, whereas it will be more difficult to detect the effects of rare variation in some other genes, as too many causal variants have been lost due to the founder effect.

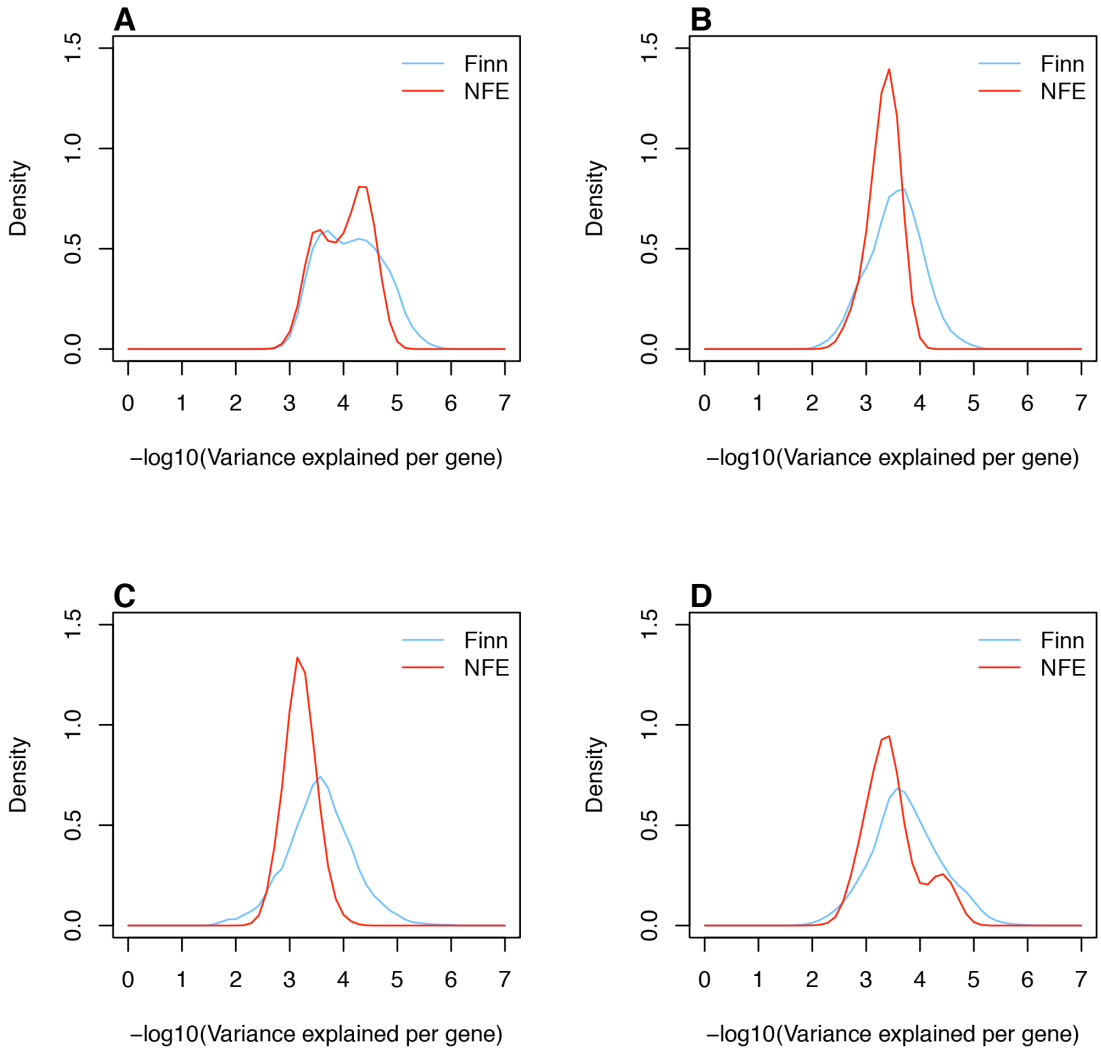

Figure 4.4: Distribution of variance explained per gene. Distribution of variance explained per gene by variants with MAF below 5% under four different disease models in either 30,000 Finns or 30,000 NFEs. (A) M1 (τ=0); (B) M2 (τ=0.5); (C) M3 (τ=1); (D) M4 (τ randomly sampled from 0, 0.5, and 1 for each effect gene).

Founder population *vs*. non-founder population in exome sequencing studies

With simulated genotype and phenotype data, we compared the power of using 30,000 NFEs and 30,000 Finns in exome sequencing studies under different disease models. We implemented five gene-based tests (T1, T5, MB, VT, SKAT-O) and single variant tests. Because we are interested in the role of lower frequency variants, all tests were run on variants with MAF below 5%. In the context of exome sequencing studies, the significance threshold for calculating power is set at $\alpha=2.5\times10^{-6}$ (after Bonferroni correction, assuming 20,000 genes in the exome).

As seen in Figure 4.5, as τ increases, so does the power from using Finns compared with using NFEs (compare panels A, B and C). Under M4, the biggest power gain in the Finns is seen among genes of which τ value is 1 (Figure S4.7). As the value of τ increases, phenotypic impacts of rare variants increase (Figure S4.4). Therefore it is more powerful to use a founder population in models where rare variation plays a more prominent role. These results are consistent with the effect of a founder event on allele frequencies – founder effects impact rare variants more than common variants.

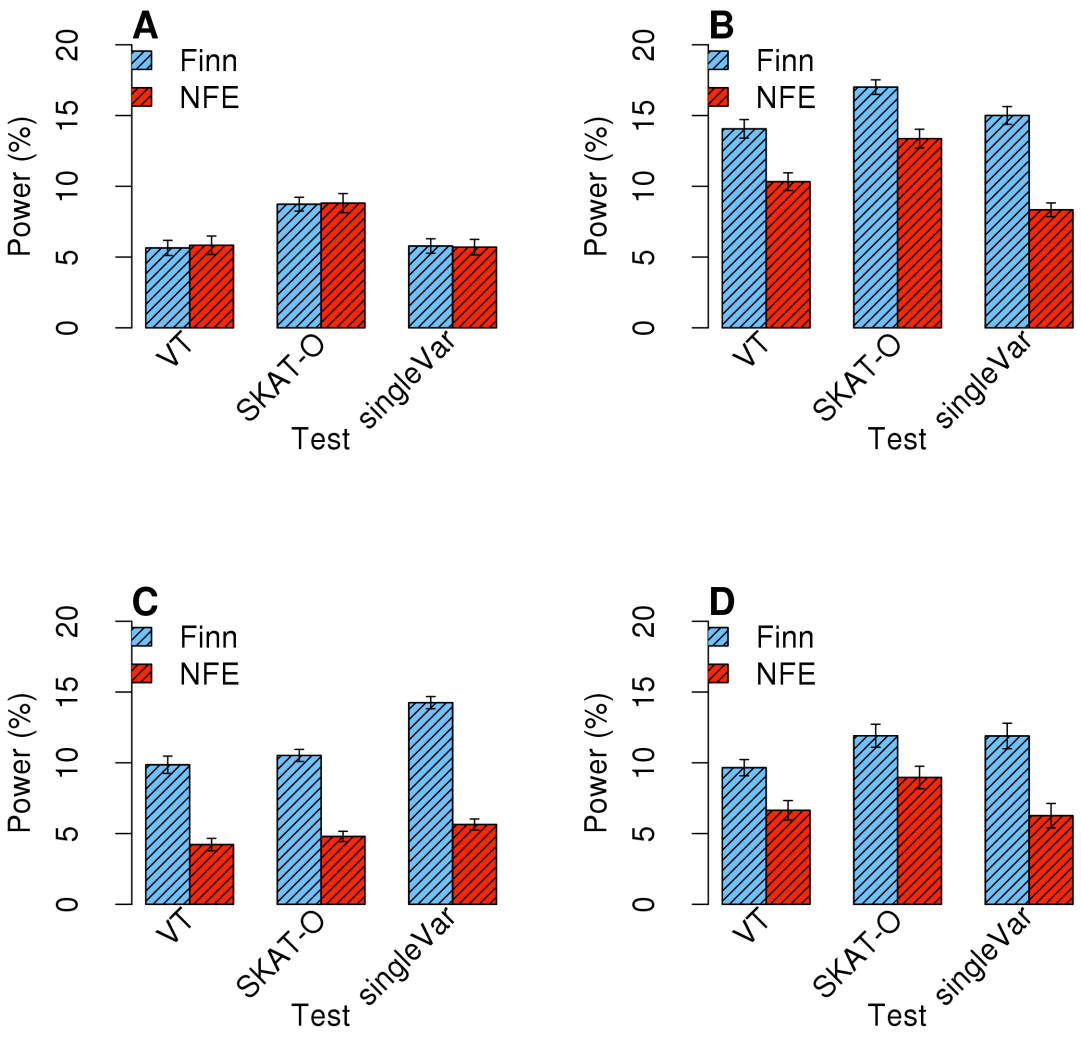

Figure 4.5: Power of exome sequencing studies in 30,000 Finns *vs.* 30,000 NFEs.
(A) M1 (τ=0); (B) M2 (τ=0.5); (C) M3 (τ=1); (D) M4 (τ randomly sampled from 0, 0.5, and 1 for each effect gene). We simulated a quantitative trait (h^2 = 80%) for which aggregated coding variation in 1,000 genes explains the total heritability. Models M1-4 were generated by varying the degree of coupling (τ) between a causal variant's phenotypic effect and the strength of purifying selection against that variant. We compared SKAT-O, variable threshold test (VT) and single variant tests (singleVar).

Understanding the excess of power in the founder population

To understand why there is an excess of power in the founder population across different disease models, we considered genes detected in one population only. We observed that genes detected only in the Finns tend to have greater variance explained per gene in the Finns *vs.* the NFEs (Figure S4.8), while the opposite is true for genes detected in the NFEs only (Figure S4.9). For genes detected in the Finns only, the cumulative allele frequency for background variants is similar between the Finns and the NFEs, but the cumulative allele frequency for causal variants is shifted upwards in the Finns (Figure S4.10). For genes detected in the NFEs only, the opposite is true (Figure S4.11).

As shown above, overall frequency rise of causal variants and thus greater variance explained for some genes could drive the excess of power in the Finns. We next tested whether reduced heterogeneity could also contribute to the power difference. We selected a set of genes for which the variance explained is closely matched between the NFEs and the Finns (Figure S4.12). As shown in Figure S4.13, the accumulated allele frequency of causal variants and background variants are similar between the two populations as well. The power gain in the Finns is retained under M3 (Figure S4.14), suggesting that reduced genetic heterogeneity alone could increase power when variance explained at a gene stays the same in the founder population. This effect is clearer when τ value is 1, as rare variants play a more prominent role.

Relative power of different association tests for rare variants in a founder population

Among the five gene-based tests we conducted, SKAT-O and VT tests perform best across a range of models in both the Finns and the NFEs (Figure 4.5, S4.15), as SKAT-O allows different variants to have different directions and magnitude of effects and VT decreases

background noise by selecting an optimal frequency threshold. The single variant test performs reasonably well under different disease models and it is particularly powerful when τ is large, especially when used in a founder population (Figure 4.5, S4.15). As τ increases, the effect sizes of rare causal mutations tend to increase, making it more powerful to test these variants individually. For a founder population like the Finns, as we have shown earlier, the allele frequency spectra are shifted away from the rarest variants, which gives extra power in testing rare variants individually.

Of note, LD has not been taken into account in both the single variant and gene-based tests. For gene-based tests, LD is generally not addressed, at least for discovery of gene-wide association signals. For single variant tests, we have operated under the assumption that the causal variants are directly assessed. Incorporating LD might further increase the power of single variant analyses, if one or more very rare variants were tagged by a single more common variant.

Exome chip studies *vs.* exome sequencing studies

Exome chip genotyping, despite being a much cheaper technology than exome sequencing, has not been rigorously assessed in terms of cost-efficiency. Here we used our simulation framework to try to address this question. We first confirmed that our simulations reproduced the expected differences in observed allele frequency spectra between exome chip and exome sequence data (Figure S4.16, compare with Figure 4.3B). We then compared the cost-efficiency of exome chip studies and exome sequencing studies under different disease models in the Finnish founder population. The cost of exome chip per sample was assumed to be about one tenth that of exome sequencing. Figure 4.6 shows that under M4, the power of SKAT-O is far greater in exome chip studies than exome sequencing studies at a fixed cost (middle versus

bottom line); in a fixed number of samples, however, genotyping arrays miss a substantial portion of causal variation detected in sequencing (top versus middle line). We also compared the two study designs in a non-founder population (Figure S4.17), under M1-3 (Figure S4.18, S4.19), as well as using different rare variant association tests (Figure S4.18), and observed similar results. Despite substantial cost-efficiency, the exome chip is underpowered to detect the contributions of certain genes simply because not enough causal variants in these genes are covered by the chip. This becomes more apparent as τ increases (Figure S4.19), because the allele frequency spectrum of causal variants shifts downwards, and the exome chip captures fewer casual variants (Figure S4.20).

As Finnish samples contributed to exome chip design, we went on to assess how much their inclusion impacts the power of the exome chip in Finns, i.e., how the exome chip would perform in a non-Finnish founder population. To address this question, we simulated a different exome chip, with no contribution of Finnish samples (replaced with an equal number of NFE samples). As expected, the power for rare variant association in Finns decreases when Finns were not used in the SNP discovery process of the exome chip. The power decrease was minimal at low sample size, reaching a difference of approximately 2% when 30,000 samples were used (Figure S4.21; the power dropped from ~10% to ~8% when Finns were not used). This suggests that the current exome chip would perform slightly less well in a non-Finnish founder population. However, the exome chip is still a far more cost-efficient strategy for such populations compared to exome sequencing, the power of which at a comparable cost (bottom line in Figure 4.6) is negligible. If it is desirable to avoid the marginal loss of power in non-Finnish founder populations of interest, one could perform exome sequencing using a representative population sample first, and supplement the exome chip with the newly

discovered variants to ensure that rare variation in that founder population is directly represented on the chip.

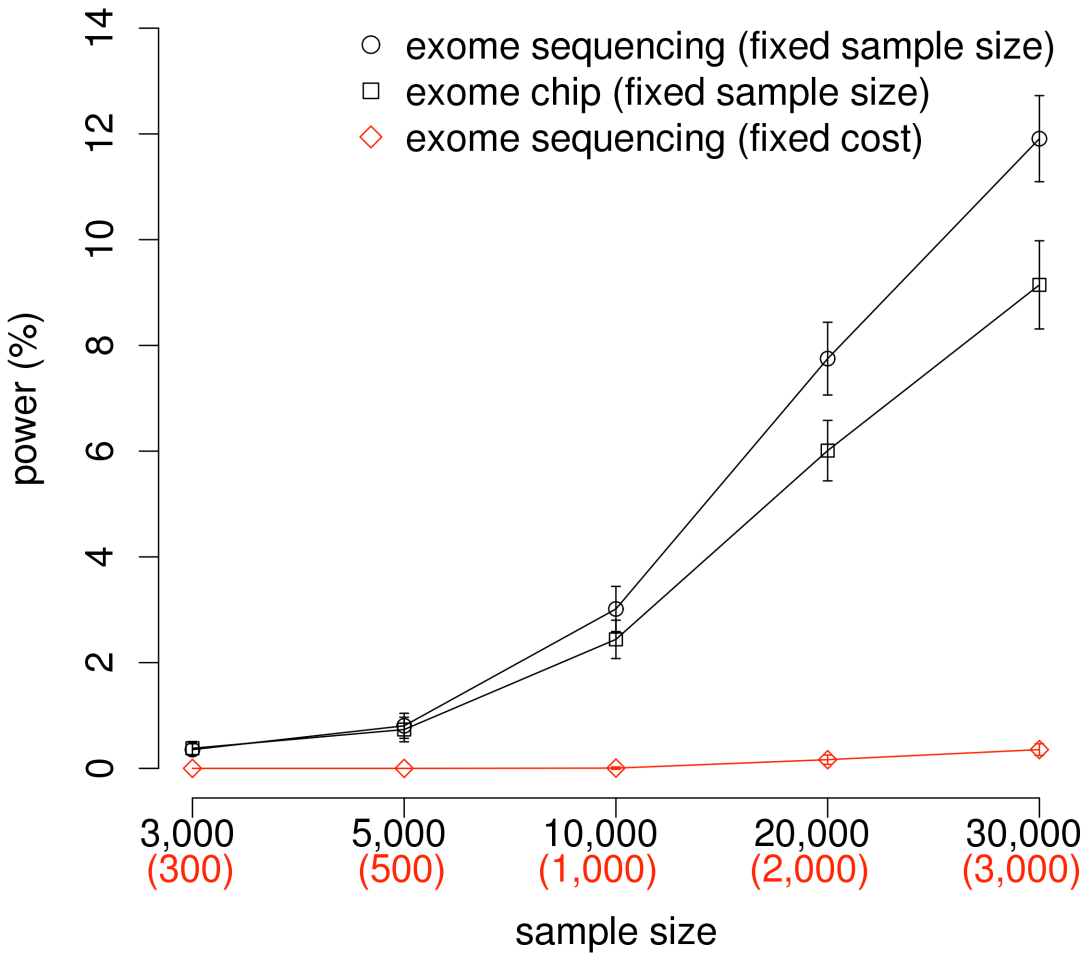

Figure 4.6: Power of exome chip study *vs.* exome sequencing study in the Finns. The comparison was done under M4 using SKAT-O test. As different genes are likely to have different pleiotropic effects and are therefore exposed to different strengths of purifying selection, M4 is generated to represent a potentially more realistic scenario. The top two lines show power comparison at a fixed sample size; the bottom two lines show power comparison at a fixed cost (and thus only a tenth of the samples were sequenced).

DISCUSSION

By using forward simulations based on empirical deep resequencing data, we showed that 1) founder populations can provide additional power, especially when the phenotypic effects of variants are tightly coupled with effects on fitness; 2) in a founder population, the single variant test, SKAT-O and VT perform best under different disease models, and the single variant test is particularly powerful when the phenotypic effects of variants are tightly coupled with effects on fitness; and 3) exome chip genotyping is currently much more cost-efficient than exome sequencing, but misses a substantial portion of causal variation in a sequencing study of the same sample size. We also suggest that more than 10,000 samples will likely be required to reach non-negligible statistical power to identify associations with low frequency variation, assuming a per-gene contribution of ~0.1% of heritability. This is consistent with recent independent estimates of required sample sizes[11,24]. We are almost certainly underestimating the required sample size, as we modeled a highly heritable trait where all heritability is explained by coding variants in 1,000 genes, while the average contribution of coding variation to heritability will likely typically be lower than 0.1% per gene.

The changes in allele frequency and decreased allelic diversity in founder populations that are caused by the bottleneck event(s) and drift can aid in detection of rare variant associations. We have shown through our simulation that the power gain in the founder population is from both increased frequency of causal variants (thus variance explained per gene) at some genes and reduced genetic heterogeneity. Founder populations typically also demonstrate a higher degree of cultural and environmental homogeneity (not modeled here), which could further increase the strength of the genetic signals. However, there are also limitations with using founder populations. First, the population size may not be large enough to

allow for the collection of sufficiently large numbers of cases. Second, rare variants may be recent in origin and hence specific to a single founder population; these are the variants that are potentially least replicable, although this concern is less relevant for gene-based "burden"-type tests where variants are aggregated. Of note, the variants might be unique to founder populations, but the finding of genes is relevant to all populations. Third, a higher rate of direct and cryptic relatedness in some founder populations could confound baseline assumptions of independence among genotypes and phenotypes and may require more specialized approaches to account for this sample structure. Fourth, there might not be enough power to detect some genes in the founder population due to loss of causal variants (Figure 4.4). Nevertheless, the increased power, particularly for single variant tests, suggests that exome chip and/or exome sequencing in all available samples from founder populations would be an efficient use of resources. Different founder populations will happen to be better powered for different genes (in each population, certain genes will gain in power while others lose power, but the genes that gain in power will vary across populations). Thus, a potentially attractive strategy for rare variant studies is to employ a diverse panel of well-powered founder populations.

We evaluated a variety of statistical tests that were developed using different assumptions about genetic architecture. We have shown that these tests are indeed sensitive to different disease models. SKAT-O and VT tests outperform the other gene-based tests across a range of different genetic architectures. It is also worth noting that single variant tests perform as well as or better than SKAT-O and VT, particularly in founder populations with decreased allelic diversity. This raises as one possible strategy to use single variant tests as a screen in founder populations and then follow up with candidate gene sequencing.

We have shown exome chip genotyping studies are currently much more cost-efficient than exome sequencing studies under a range of genetic models. In a fixed number of samples, however, exome chip genotyping studies miss a substantial portion of causal variation that could be detected by sequencing. Continued sharp drops in the cost of sequencing and/or targeted sequencing to follow up initial results might enable better-powered and more cost-efficient exome sequencing or whole genome sequencing studies. Given the requirement for large sample sizes, the ability to combine studies for example in meta-analyses will be critical for a new wave of discoveries. Of note, as reference panels for imputation become larger and represent more populations, imputation of rare variants into samples with existing genotype data is another likely complementary approach for future studies.

Our study has a number of limitations. We have taken a forward-in-time approach for simulating population sequence data, which has substantial advantage in terms of being able to model different genetic architectures and demographic parameters, but this approach comes with the cost of requiring greater computational resources. Because of this limitation, as well as the complexity of the demographic models, we did not do a complete search through the entire parameter space for the best-fitting demographic model. Another limitation with our simulation is that the limited sample size of the empirical data provided an incomplete view of rare variants in the population, so our simulations may not be completely accurate at very low allele frequencies. Moreover, as suggested by Casals et al.[25], there might be a relaxation of selection in the founder population, which we have not considered in our simulation. It is also worth noting that our empirical Finnish samples are from all across Finland (Table S4.2) and therefore our model for simulating the Finns ignored the demographic heterogeneity within Finland. As deeper and richer human genetic data becomes available, the models can be calibrated and improved.

Last but not least, our study does not explore the effects of properties such as gene size or mutation rate on power, nor does it characterize power of rare variant tests at non-coding loci, where causal variant frequencies and effect sizes may be different.

In summary, our study has highlighted the usefulness of understanding the population-genetic properties of a study population to explore a range of genetic models and recognize the features and limitations of different association study designs in that population. As the field of human genetics moves forward to explore new and expanded sources of variation, such models offer a context with which to interpret the data and to plan future studies for gene discovery. With current approaches focused on rare variation, our work suggests that founder populations such as Finland can play an important role in genetic studies.

MATERIALS AND METHODS

Empirical Exome Sequencing Data

We used whole exome sequenced samples from GoT2D (Genetics of Type 2 Diabetes) Project. In total, 2850 European type 2 diabetes cases and controls from four cohorts (DGI, FUSION, GoT2D-UK, KORA) were whole exome sequenced at ~40X. Exome target capture was performed with the Agilent SureSelect Human All Exon hybrid selection kit and sequence obtained on HiSeq. Subsequent alignment and allele calling used Burrows-Wheeler Aligner (BWA)[26] and Genome Analysis Tookit (GATK)[27]. We kept samples from GoT2D-UK and KORA as the NFE population and samples from FUSION (**Table S1**) as the Finnish population for our analyses. We excluded SNPs with any missing data in any individual, SNPs with Hardy-Weinberg equilibrium (HWE) $P < 10^{-5}$, and all nonautosomal SNPs. We carried out multidimensional scaling (MDS) to identify population outliers (Figure S4.22). We filtered out

relatives, for whom the estimated genome-wide identity-by-descent (IBD) proportion to alleles shared was >0.10. We also excluded individuals with inbreeding coefficient >0.05 or <-0.05. We estimated IBD sharing using PLINK's '--genome' option[28] and estimated inbreeding coefficients using PLINK's '--het' option. All analyses were carried out on an LD-pruned set of SNPs obtained by using the PLINK option '--indep', which recursively removes SNPs within a sliding window. The parameters for --indep are: window size in SNPs (50), the number of SNPs to shift the window at each step (5), and the variance inflation factor (VIF) threshold (1.8). The final dataset included 843 Finnish samples and 820 NFE samples.

Simulation of Exome Sequencing Data

Exome sequencing data were simulated using ForSim, a forward evolutionary simulation tool.[20] The average gene coding length was set as 1500bp. We used a mutation rate per site of $2x10^{-8}$ [29–31] and a uniform locus-wide recombination rate of 2Mb/cM as in previous report[22]. We modeled the distribution of selection coefficients for de novo missense mutations by a gamma distribution[32] (as in previous reports[32,33], we assume that ~20% of missense sites are neutrally evolving).

For modeling the NFEs, we used a conventional four-parameter model of the history of the European population with long-term constant size followed by a bottleneck and then by an exponential expansion (Figure 4.2)[21]. The four parameters used were: (1) long-term ancestral effective population size; (2) bottleneck population size; (3) duration of exponential growth in generations; and (4) recent effective population size. We adapted parameters from a recent simulation that generated representative sequence data for European populations[22]. (see Figure 4.2 for final parameter values)

We modeled the Finns as a founder population established by a small number of NFEs. The founding event was followed by a slow growth phase and a more recent fast growth phase (Figure 4.2). The demographic history parameters were fitted by comparing to empirical exome sequencing data. P(data|model) was calculated and used for model-fitting (see Supplemental Method for further discussion). We tested two other models, neither of which agreed with the empirical observations as well as our current model (Table S4.3).

Simulation of Exome Chip Data

Exome chip data were generated based on simulated exome sequencing data. The process resembles that of the actual exome chip design (see URL for a description of SNP content and selection strategies). ~12,000 simulated exomes across 16 cohorts were pooled together. The cohorts were matched by ancestry and sample size with the real cohorts (except that non-European samples were substituted by NFE samples). Only missense variants observed three or more times in at least two datasets were selected. ~90% of the selected SNPs passed design and were used for the simulated chip. For simulating the exome chip without contribution of Finnish samples, all procedures are the same except that Finnish samples are replaced with an equal number of NFE samples.

Simulation of Phenotypic Variation

We simulated a quantitative trait (QT) with a target size of 1,000 genes and the heritability of 80%. For efficiency, we modeled the heritability as completely explained by coding variants and a large target size of 1,000 genes; power will scale with total heritability, fraction of heritability explained by coding variation, and inversely with target size. We modeled

additive genetic effects as well as environmental effects. We did not consider non-additive effects (dominant, recessive, epistatic, gene-environment interactions) in our simulations. The joint allele frequency spectrum of both the Finns and the NFEs is used when calculating heritability. The effect sizes are scaled so as to cap heritability at 80%. More specifically, variance explained by a variant is calculated as $2*MAF*(1-MAF)*p^2$, where p is phenotypic effect. We sum this up over all variants and to cap heritability (additive genetic variance) at 80%, effect sizes of all variants are adjusted by a uniform factor.

We assume neutral missense variants have no effect on phenotype. For assigning effect sizes of causal variants (non-neutral missense variants), we implemented a range of possible mappings between a variant's selection coefficient s (we modeled the distribution of selection coefficients for de novo missense mutations by a gamma distribution, so s is known for every variant in our simulated data) and its effect on phenotype (p). We model these mappings as: $p=s^{\tau}*(1+\varepsilon)$ as suggested by Eyre-Walker et al[23]. Here, τ is the degree of coupling between p and s; ε is a normally distributed random noise parameter. In the case of common diseases of post-reproductive onset, the role of natural selection on causal variants is not yet clear. Therefore we tested a range of scenarios: M1 (τ=0), M2 (τ=0.5), M3 (τ=1), and M4 (τ randomly chosen with equal probability among 0, 0.5 and 1 for each effect gene).

For determining the direction of effect of causal variants on QT, we further assumed that, in each trait-affecting gene, 0-20% of the causal variants influence the QT in the opposite direction from the remaining causal variants. This assumption is based on two different arguments: (1) the vast majority of de novo amino acid mutations with a measurable effect reduce protein activity and gain-of-function mutations are much less frequent and are restricted

to specific residues or domains; (2) some genes, like *APOB* (MIM 107730) and *PCSK9* (MIM 607786)[18,34], clearly illustrate a mixture of variants that affect QTs in both directions.

Association Tests and Power Analysis

We conducted five different gene-based tests on simulated data. Four burden tests – VT[14], T1, T5, and MB[16] -- were performed by running SCORE-Seq (See URL for details). VT stands for variable-threshold test; T1 and T5 are fixed threshold tests and pertain to the threshold of 1% and 5% respectively; MB stands for Madsen and Browning. The mutation information is aggregated across multiple variant sites of a gene through a weighted linear combination and then related to the phenotype of interest through appropriate regression models. The weights can be constant (T1, T5, VT) or dependent on allele frequencies (MB). The allele-frequency threshold can be fixed (T1, T5, MB) or variable (VT). We also performed the unified optimal test SKAT-O using default weights[35]. SKAT-O is a data-adaptive test that includes both burden tests and SKAT[17] as special cases. Single-variant tests were carried out using PLINK's '--linear' option. We limited our analysis to variants with minor allele frequency below 5%.

The exome-wide significance threshold for gene-based tests is set at $\alpha=2.5\times10^{-6}$ (after bonferroni correction, assuming 20,000 genes in the exome). Power is defined as the number of effect genes reaching genome-wide significance divided by the target size, which is 1000. The exome-wide significance threshold for single variant test is 0.05 divided by the number of variants tested (varying with sample size, excluding singletons and doubletons). The power of the single variant test is defined as the number of effect genes harboring genome-wide significant variant(s) divided by 1000.

ACKNOWLEDGEMENTS

We gratefully acknowledge B. Lambert and K. Weiss (authors of the simulation tool ForSim) for helpful technical assistance. Without their software, this work would not have been possible. We also thank M. McCarthy and M. Boehnke for discussion and insightful critiques and M. Lin for helping with the preparation of figures. This work was supported by grant from US National Institute of Health (NIH; award 2R01DK075787 to J.N.H.). V.A. is supported by NIH Training grants T32GM007753 and T32GM008313. J.F. is supported in part by NIH Training grant T32GM007748-33 as well as by funding from Pfizer. C.W.K.C. is supported by NIH NRSA Postdoctoral Fellowship F32GM106656. The GoT2D Study is supported by grant 1RC2DK088389-01 from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK).

Web Resources

The URLs for data presented herein are as follows:

Online Mendelian Inheritance in Man (OMIM), http://www.omim.org/

Exome array design, http://genome.sph.umich.edu/wiki/Exome_Chip_Design

PLINK, http://pngu.mgh.harvard.edu/~purcell/plink/

SCORE-Seq: Score-Type Tests for Detecting Disease Associations With Rare Variants in Sequencing Studies, http://www.bios.unc.edu/~lin/software/SCORE-Seq/

CONSORTIUM

Broad Institute: Jason Flannick, Alisa Manning, Christopher Hartl, Vineeta Agarwala, Pierre Fontanillas, Todd Green, Eric Banks, Mark DePristo, Ryan Poplin, Khalid Shakir, Timothy Fennell, Jacquelyn Murphy, Noël Burtt, Stacey Gabriel, David Altshuler

University of Michigan / FUSION: Christian Fuchsberger, Hyun Min Kang, Xueling Sim, Clement Ma, Adam Locke, Thomas Blackwell, Anne Jackson, Tanya Teslovich, Heather Stringham, Peter Chines, Phoenix Kwan, Jeroen Huyghe, Adrian Tan, Goo Jun, Michael Stitzel, Richard N. Bergman, Lori Bonnycastle, Jaakko Tuomilehto, Francis S. Collins, Laura Scott, Karen Mohlke, Gonçalo Abecasis, Michael Boehnke
Helmholtz München/KORA: Tim Strom, Christian Gieger, Martina Müller-Nurasyid, Harald Grallert, Jennifer Kriebel, Janina Ried, Martin Hrabé de Angelis, Cornelia Huth, Christa Meisinger, Annette Peters, Wolfgang Rathmann, Konstantin Strauch, Thomas Meitinger

Lund University: Jasmina Kravic, Claes Ladenvall, Tiinamaija Toumi, Bo Isomaa, Leif Groop

Univ of Oxford & Wellcome Trust: Kyle Gaulton, Loukas Moutsianas, Manny Rivas, Richard Pearson, Anubha Mahajan, Inga Prokopenko, Ashish Kumar, John Perry, Jeff Chen, Bryan Howie (Chicago), Martijn van de Bunt, Kerrin Small (Kings), Cecilia Lindgren, Gerton Lunter, Neil Robertson, Will Rayner, Andrew Morris, David Buck, Andrew Hattersley (Exeter), Tim Spector (Kings), Gil McVean, Tim Frayling (Exeter), Peter Donnelly, Mark McCarthy

REFERENCES

1. Peltonen, L., Jalanko, A., and Varilo, T. (1999). Molecular genetics of the Finnish disease heritage. Hum. Mol. Genet. *8*, 1913–1923.

2. Nevanlinna, H.R. (1972). The Finnish population structure. A genetic and genealogical study. Hereditas *71*, 195–236.

3. De la Chapelle, A. (1993). Disease gene mapping in isolated human populations: the example of Finland. J. Med. Genet. *30*, 857–865.

4. Lahermo, P., Sajantila, A., Sistonen, P., Lukka, M., Aula, P., Peltonen, L., and Savontaus, M.L. (1996). The genetic relationship between the Finns and the Finnish Saami (Lapps): analysis of nuclear DNA and mtDNA. Am. J. Hum. Genet. *58*, 1309–1322.

5. De la Chapelle, A., and Wright, F.A. (1998). Linkage disequilibrium mapping in isolated populations: the example of Finland revisited. Proc. Natl. Acad. Sci. U. S. A. *95*, 12416–12423.

6. Kittles, R.A., Perola, M., Peltonen, L., Bergen, A.W., Aragon, R.A., Virkkunen, M., Linnoila, M., Goldman, D., and Long, J.C. (1998). Dual origins of Finns revealed by Y chromosome haplotype variation. Am. J. Hum. Genet. *62*, 1171–1179.

7. Peltonen, L., Pekkarinen, P., and Aaltonen, J. (1995). Messages from an isolate: lessons from the Finnish gene pool. Biol. Chem. Hoppe. Seyler. *376*, 697–704.

8. Hedman, M., Pimenoff, V., Lukka, M., Sistonen, P., and Sajantila, A. (2004). Analysis of 16 Y STR loci in the Finnish population reveals a local reduction in the diversity of male lineages. Forensic Sci. Int. *142*, 37–43.

9. Sajantila, A., Salem, A.H., Savolainen, P., Bauer, K., Gierig, C., and Pääbo, S. (1996). Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population. Proc. Natl. Acad. Sci. U. S. A. *93*, 12035–12039.

10. Service, S., DeYoung, J., Karayiorgou, M., Roos, J.L., Pretorious, H., Bedoya, G., Ospina, J., Ruiz-Linares, A., Macedo, A., Palha, J.A., et al. (2006). Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. Nat. Genet. *38*, 556–560.

11. Zuk, O., Schaffner, S.F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M.J., Neale, B.M., Sunyaev, S.R., and Lander, E.S. (2014). Searching for missing heritability: Designing rare variant association studies. Proc. Natl. Acad. Sci. *111*, E455–64.

12. Jonsson, T., Atwal, J.K., Steinberg, S., Snaedal, J., Jonsson, P. V, Bjornsson, S., Stefansson, H., Sulem, P., Gudbjartsson, D., Maloney, J., et al. (2012). A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. Nature *488*, 96–99.

13. Asimit, J., and Zeggini, E. (2010). Rare variant association analysis methods for complex traits. Annu. Rev. Genet. *44*, 293–308.

14. Price, A.L., Kryukov, G. V, de Bakker, P.I.W., Purcell, S.M., Staples, J., Wei, L.-J., and Sunyaev, S.R. (2010). Pooled association tests for rare variants in exon-resequencing studies. Am. J. Hum. Genet. *86*, 832–838.

15. Li, B., and Leal, S.M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet. *83*, 311–321.

16. Madsen, B.E., and Browning, S.R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. *5*, e1000384.

17. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. *89*, 82–93.

18. Neale, B.M., Rivas, M.A., Voight, B.F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S.M., Roeder, K., and Daly, M.J. (2011). Testing for an unusual distribution of rare variants. PLoS Genet. *7*, e1001322.

19. Pritchard, J.K., and Cox, N.J. (2002). The allelic architecture of human disease genes: common disease-common variant...or not? Hum. Mol. Genet. *11*, 2417–2423.

20. Lambert, B.W., Terwilliger, J.D., and Weiss, K.M. (2008). ForSim: a tool for exploring the genetic architecture of complex traits with controlled truth. Bioinformatics *24*, 1821–1822.

21. Adams, A.M., and Hudson, R.R. (2004). Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. Genetics *168*, 1699–1712.

22. Agarwala, V., Flannick, J., Sunyaev, S., and Altshuler, D. (2013). Evaluating empirical bounds on complex disease genetic architecture. Nat. Genet.

23. Eyre-Walker, A. (2010). Evolution in health and medicine Sackler colloquium: Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. Proc. Natl. Acad. Sci. U. S. A. *107 Suppl* , 1752–1756.

24. Kiezun, A., Garimella, K., Do, R., Stitziel, N.O., Neale, B.M., McLaren, P.J., Gupta, N., Sklar, P., Sullivan, P.F., Moran, J.L., et al. (2012). Exome sequencing and the genetic basis of complex traits. Nat. Genet. *44*, 623–630.

25. Casals, F., Hodgkinson, A., Hussin, J., Idaghdour, Y., Bruat, V., de Maillard, T., Grenier, J.-C., Gbeha, E., Hamdan, F.F., Girard, S., et al. (2013). Whole-Exome Sequencing Reveals a Rapid Change in the Frequency of Rare Functional Variants in a Founding Population of Humans. PLoS Genet. *9*, e1003815.

26. Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595.

27. DePristo, M.A., Banks, E., Poplin, R., Garimella, K. V, Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

28. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

29. Nachman, M.W., and Crowell, S.L. (2000). Estimate of the mutation rate per nucleotide in humans. Genetics *156*, 297–304.

30. Kondrashov, A.S. (2003). Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. Hum. Mutat. *21*, 12–27.

31. Sun, J.X., Helgason, A., Masson, G., Ebenesersdóttir, S.S., Li, H., Mallick, S., Gnerre, S., Patterson, N., Kong, A., Reich, D., et al. (2012). A direct characterization of human mutation based on microsatellites. Nat. Genet. *44*, 1161–1165.

32. Kryukov, G. V, Shpunt, A., Stamatoyannopoulos, J.A., and Sunyaev, S.R. (2009). Power of deep, all-exon resequencing for discovery of human trait genes. Proc. Natl. Acad. Sci. U. S. A. *106*, 3871–3876.

33. Ahituv, N., Kavaslar, N., Schackwitz, W., Ustaszewska, A., Martin, J., Hebert, S., Doelle, H., Ersoy, B., Kryukov, G., Schmidt, S., et al. (2007). Medical sequencing at the extremes of human body mass. Am. J. Hum. Genet. *80*, 779–791.

34. Kotowski, I.K., Pertsemlidis, A., Luke, A., Cooper, R.S., Vega, G.L., Cohen, J.C., and Hobbs, H.H. (2006). A spectrum of PCSK9 alleles contributes to plasma levels of low-density lipoprotein cholesterol. Am. J. Hum. Genet. *78*, 410–422.

35. Lee, S., Wu, M.C., and Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. Biostatistics *13*, 762–775.

Chapter 5

Concluding Remarks

THE OVERVIEW

While some traits may be under positive or balancing selection, other traits and diseases are likely under negative selection because of their deleterious effects to the fitness of the patient. Even for putatively neutral traits, mutations could have pleiotropic effects on fitness and thus be selected against. Under the selective pressure, causal variants would be kept at a low population frequency. Therefore, rare variants at multiple loci may influence the complex disease risk[1].

Despite the potential importance of rare variants, testing for their association with diseases and traits is likely to be challenging[2,3]. Rare variants are individually of low frequencies and well-powered studies require large sample sizes. Unfortunately, performing whole-genome or whole-exome sequencing for large cohorts is still very expensive. This dissertation presented approaches in response to these challenges – we developed a cost-efficient pooled sequencing scheme for follow-up studies of candidate genes, and proposed a simulation framework for evaluating various designs of rare variant association studies (RVAS).

In this concluding chapter, I first list the major findings of each of the studies presented in this dissertation, followed by a discussion of the broader implications highlighted by these studies. As most issues directly relevant to each study have already been presented in the discussion section of each chapter, this section will be kept relatively brief. Finally, this chapter will conclude with a more general discussion on the potential outcome of RVAS.

MAJOR FINDINGS

Chapter 2

- We developed a cost-efficient pooled sequencing scheme with well-controlled false-positive and false-negative rates.

- We identified three cases with known pathogenic variants in PTPN11, highlighting the possibility that Noonan syndrome may be an underdiagnosed cause of short stature.
- The vas majority of HGMD-reported disease-causing dominant mutations did not manifest with the associated clinical phenotype in our cohort, suggesting that the classification of these variants as pathogenic is probably erroneous.
- We reported a new frameshift mutation in IGF1R and demonstrate its pathogenicity by functional studies.

Chapter 3

- We identified heterozygous NPR2 mutations in ~2% of short stature patients; family analysis demonstrated segregation of these variants with the short stature phenotype.
- We screened for NPR2 mutations in two cohorts of samples at the extremes of height distribution in population-based cohorts.
- Heterozygous NPR2 mutations in NPR2 could be an important cause of nonsyndromic familial short stature.

Chapter 4

- We have developed a population genetics framework to assess the impact of the Finnish population history on genetic studies of rarer variation; more generally, our simulation approach provides a framework to evaluate association study designs in different study populations.
- We demonstrate that power for rare variant association tests is higher in the Finnish population, especially when variants' phenotypic effects are tightly coupled with

fitness effects.

- SKAT-O, VT, and single variant tests are more powerful than other rare variant methods in a founder population.
- At a fixed cost, genotyping strategies have far greater power than sequencing; in a fixed number of samples, however, genotyping arrays miss a substantial portion of causal variation detected in sequencing.

BROADER IMPLICATION OF THE STUDIES

Ongoing RVAS attempts a search for genes harboring multiple rare variants collectively associated with complex traits. The power of such studies depends on three key quantities: the combined allele frequency of the tested alleles, the excess relative risk of disease conferred by alleles in the class, and sample size. To maximize the power of an association study, we want all these quantities, to be as large as possible. Unfortunately, these goals sometimes pull in opposite directions. For example, expanding the alleles under study increases combined allele frequency but dilutes the excess relative risk[4].

The balance between combined allele frequency and the excess relative risk

The association signal is provided by pathogenic variants, where as benign alleles are a source of noise masking the association signal. Ideally, one would aggregate only pathogenic alleles and ignore benign alleles. Unfortunately, one cannot perfectly distinguish the former from the latter. To enrich for harmful alleles, RVAS typically focuses on nonsynonymous variants in protein-coding regions, or focuses on variants with frequency below a specified threshold. Even with these limitations, the resulting variants remain a mixture of pathogenic and benign alleles.

Several potential directions to optimize the balance between combined allele frequency and the excess relative risk can be pursued.

Functional characterization of human allelic variants

The ability to discriminate between pathogenic and benign alleles would dramatically increase the potential of sequencing studies focusing on rare variants in complex traits. Several studies have demonstrated that highlighting functional variants using experimental[5,6] or computational[7–9] approaches increase the power of these studies.

Medical genetics is interested in finding "pathogenic" mutations that causally influence traits of interest. Population genetics focuses on "deleterious" alleles that are under purifying selection. Functional analysis is focused on the "damaging" effect on molecular function. The rationale for current approaches that infer functional significance of human allelic variants is that the effects on phenotypes correlate with the effects on fitness and are mediated by the effects on molecular function. Yet it is possible that most of human alleles under purifying selection have no detectable effects on medically relevant phenotypes in current environment, and damaging alleles may be neutral or beneficial rather than deleterious[10]. Below we review major prediction methods and discuss the limitations or future direction for each.

1. Allele frequency as a proxy

The analysis of allele frequency in unaffected controls has been used in many studies to enrich for pathogenic variants. This approach has been dramatically facilitated by large-scale sequencing efforts such as 1000 Genomes Project[11] and Exome Sequencing Project (ESP)[12]. It is easy to infer that the variant is benign (or, at least, not of high penetrance) if it is seen at appreciable frequency in healthy controls, yet the sole observation of its absence in multiple

controls is insufficient to convincingly imply pathogenicity. Other limitations include imbalanced number of cases and controls in some studies and no phenotype information for sequenced individuals in public datasets such as 1000 Genomes and ESP.

2. Experimental evidence

Direct experimental functional analysis is a highly convincing method to test the effect of human allelic variants, which includes the analysis of protein expression and localization, in vitro functional assays and genetic manipulation on model organisms. The limitations with direct experimental methods are that: they are highly laborious and not feasible, at least currently, at the whole-exome scale; it is not always easy to find functional assays that are informative about the human condition.

3. Computational predictions

Computational programs, such as PlolyPhen-2[13], SIFT[14], or Mutation Taster[15], offer predictions of whether a mutation is likely to be damaging. Prediction methods mostly rely on two fundamental observations. First, the analysis of phylogenetic information in the form of multiple sequence alignment is a powerful source of information about the spectrum of residues allowed at particular positions of the protein of interest. Second, mapping mutations on protein 3D structure may provide key insights into the functional mechanisms, if the structure has been resolved for the protein of interest or its close homolog[10].

At this time, the accuracy of computational predictions is ~75-80%, which is less informative than direct experimental evidence[10]. However, given that computational methods do not involve any additional labor and cost and can be applied to any gene, they will likely continue to be used widely in the future. On the one hand, as more protein sequences and structures accompanied by training data (known disease-causing mutations and neutral

polymorphisms) are available, the classification accuracy will improve. On the other hand, the accuracy can be potentially improved if the scope of the methods would be narrower, so they would be specifically focused on a single phenotype and a group of genes involved in this phenotype.

Extending RVAS to noncoding regions

By extending RVAS to noncoding regions, more pathogenic variants can be included in the association tests. A major challenge, though, is how to select the genomic regions across which to aggregate variants. To perform RVAS with reasonable sensitivity in noncoding regions, it will be important to have fairly precise knowledge of the functionally important regulatory sequences related to each human gene to aggregate them together.

At present, without such knowledge, it is more cost-efficient to perform whole-exome, rather than whole-genome sequencing studies. This approach will maximize the number of samples that can be analyzed for coding regions, where the power is currently vastly greater, where the effect sizes are expected to be larger, and where the discoveries are likely to be more immediately actionable.

Using founder populations

Another approach is to use populations in which the combined allele frequency of casual variants for some genes happens to be much larger than other populations. As we have shown in Chapter 4, the distribution of combined allele frequency of causal variants is wider in the Finns compared to the non-Finnish Europeans. Studying recently bottlenecked populations such as Finland will make it easier to discover some disease-associated genes, although the power for

detecting some other genes will be decreased. Studying multiple founder populations may be a powerful strategy. The discoveries might be incomplete, but they may prove valuable by providing the initial insights into disease pathogenesis and architecture.

Enabling studies of larger sample size

Extrapolation of effect sizes and frequencies from published candidate gene studies shows that thousands of individuals are required for whole exome sequencing studies to reach acceptable statistical power[16]. In Chapter 4, we showed that more than 10,000 samples will likely be required to reach non-negligible statistical power to identify associations with low frequency variation (assuming a per-gene contribution of ~0.1% of heritability). This is consistent with recent independent estimates of required sample sizes based on population genetics simulations[3,4].

The cost of sequencing a human genome is dropping rapidly, due to the continual development of new, faster, cheaper DNA sequencing technologies such as next-generation DNA sequencing. Prices are expected to drop further over the next few years, with new DNA sequencing methods currently under development, such as nanopore DNA sequencing[17–19] and microscopy-based techniques[20]. However, at this time, whole-genome sequencing or even whole-exome sequencing is too expensive to perform for large sample sizes. A few techniques/designs could fill the niche between genome-wide CVAS and whole-genome sequencing.

Exome chip

The exome chips provide an economical method to assay a large number of coding variants

and investigate the role of rare DNA variants in causing disease. As we have shown in Chapter 4, at a fixed cost, exome chip has far greater power than exome sequencing, though in a fixed number of samples, exome chip misses a substantial portion of genetic signals detected in sequencing.

A limitation of exome chips is that they will miss a significant fraction (~15-20%) of variants whose genomic context is incompatible with array-based genotyping, variants highly specific to non-European populations, as well as the rarest variants in any population[21]. While these exome chip studies will only provide an imperfect approximation to the results of sequencing studies, they will provide a preview of the discoveries that will be possible when exome sequencing is performed on hundreds of thousands of samples. Given that over a million and a half exome chips have been sold, we are expecting to see interesting findings coming out of these studies especially when methods such as meta-analysis is applied to combine studies. If it turns out that large exome chip studies don't provide a lot of insight, there are two possible explanations: either coding variants have much smaller effect sizes than expected; or it's really the rarest variants missed by exome chip that are important.

Imputation

When a very large number of individuals with both exome sequence data and genome-wide genotype data are available, statistical imputation can also be a fast and economical strategy for extending sample sizes. Currently, sufficiently large reference panels that can support imputation of very rare variants are not available for most cosmopolitan populations. However, several examples of the success of this approach exist, mainly from the isolated population of Iceland. There, relatively limited genetic diversity, a panel of sequenced Icelanders, and the availability

of a large collection of genotyped individuals have enabled recent discoveries using imputation. MYH6 L721W (minor allele frequency 0.4%) was evaluated in 38,000 individuals and associated with risk for sick sinus syndrome[22] and APP A673T (minor allele frequency 0.1%) was evaluated in 71,000 individuals and associated with the risk for Alzheimer's disease[23].

Target enrichment and sample pooling

One popular efficient design for Genome-wide CVAS is the two-stage design. The first stage employs a whole-genome genotyping platform and tests all available markers for association with the disease, while the second stage uses a custom genotyping platform to follow up those markers exhibiting sufficiently strong association with the disease in the first stage. Two-stage designs gain their efficiency by excluding markers for further testing that show little evidence of association in the first stage[24]. Genome-wide RVAS will likely continue using two-stage designs, with exome sequencing or exome array genotyping used in the first stage. A range of approaches are available for follow-up in the second stage, ranging from genotype imputation to targeted genotyping or targeted sequencing. When targeted genotyping and imputation are not possible or when the association signal is driven by a burden of rare mutations, it will be necessary to undertake targeted sequencing of genes prioritized in the first stage.

Target enrichment can be a highly effective way of reducing sequencing costs and saving sequencing time. Conversely, target enrichment increases sample preparation cost and time. Assuming that the throughput of sequencing runs and our ability to analyze large whole genome sequencing datasets both continue to increase, and the cost per base of sequence continue to decrease, there will come a point at which it is no longer economical to perform target enrichment of individual samples, compared to whole genome sequencing[25].

The cost of performing target enrichment can be reduced by pooling samples before enrichment. In Chapter 2, we demonstrated that our pooled sequencing scheme reduces cost considerably and is well suited for follow-up studies on candidate genes. Several potential directions to further improve or extend the use of the approach can be pursued. First, technology development is needed for methods more readily scalable for different target sizes and different sample sizes. Second, downstream analysis methods for pooled sequencing need to be further improved. Existing methods for aggregate rare variant statistics are not designed to work well with pooled sequencing data, which lose individual genotype information and are more prone to errors compared to high-coverage individual sequencing.

OUTLOOK FOR RVAS

The combined contribution of multiple rare loci to the population-level genetic variance remains an open question because association studies that focus on rare variants remain underpowered. The few population-based common disease exome-sequencing studies published to date, have not been successful in finding individual genes showing significant enrichment[26,27]. These current findings are likely to foreshadow the definitive identification of individual genes in larger cohorts, following the trajectory of genome-wide CVAS.

Although effect sizes for rare variants may be larger than for common variants, large effect sizes or odds ratios do not equate to a large contribution to the variance explained at the population level. Recent sequencing studies identified enrichment of rare variants, but this early work suggests a large polygenic burden of rare coding variants, which alone may not account for the unexplained variation[28,29].

Methods for detecting contribution of common variants to the missing heritability have been described previously. Purcell et al.[30] developed the concept of a polygenic score by combining

the effects of multiple common variants that are modestly associated with schizophrenia. They showed that the score is predictive of schizophrenia in an independent cohort, thus indicating that there is polygenic signal from many yet-to-be-detected common variants in schizophrenia. Yang et al.[31] adopted a different approach by estimating the proportion of variance for human height explained by hundreds of thousands of common variants with a linear-model analysis. They found that at least 45% of the variance can be accounted for by common variants, indicating that there are many common variants associated with height that have yet to be discovered.

More recently, methods have been designed to detect the signal of polygenic inheritance from low-frequency variants. Chan and colleagues showed that there is more power to detect risk variants than to detect protective variants, resulting in an increase in the ratio of detected risk to protective variants. Such an excess can also occur if risk variants are present and kept at lower frequencies because of negative selection. They tested the method on published GWAS results and observed a strong signal in some diseases (schizophrenia and type 2 diabetes) but not others[32].

Agarwala et al. developed a population genetics framework to directly simulate, in large populations, a wide space of genetic architecture. Each hypothesis about genetic architecture was then quantitatively evaluated against cumulative results of empirical studies already performed. Whereas extreme models are excluded by the combination of epidemiology, linkage and genome-wide association studies, many models remain consistent, including those where rare variants explain either little (<25%) or most (>80%) of type 2 diabetes heritability[33].

Overall, the genetic architecture will be different from one complex trait to another. Even within one trait, there will be heterogeneity in phenotypic contributions across loci. Mutation rate,

overall phenotypic contribution, coupling between phenotypic effects and fitness effects and allelic spectrum of causal variants are all likely to vary across loci. Thus, outcomes of RVAS will be different for different traits. Ongoing sequencing and genotyping studies will further constrain the space of possible architectures, but very large sample sizes will be required to localize most of the heritability underlying complex traits. This is not an either-or debate, and advocating a focus on solely rare or common variants will not be a productive way forward.

REFERENCES

1. Zwick, M.E., Cutler, D.J., and Chakravarti, A. (2000). Patterns of genetic variation in Mendelian and complex traits. Annu. Rev. Genomics Hum. Genet. *1*, 387–407.

2. Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science *337*, 64–69.

3. Kryukov, G. V, Shpunt, A., Stamatoyannopoulos, J.A., and Sunyaev, S.R. (2009). Power of deep, all-exon resequencing for discovery of human trait genes. Proc. Natl. Acad. Sci. U. S. A. *106*, 3871–3876.

4. Zuk, O., Schaffner, S.F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M.J., Neale, B.M., Sunyaev, S.R., and Lander, E.S. (2014). Searching for missing heritability: Designing rare variant association studies. Proc. Natl. Acad. Sci. *111*, E455–64.

5. Bonnefond, A., Clément, N., Fawcett, K., Yengo, L., Vaillant, E., Guillaume, J.-L., Dechaume, A., Payne, F., Roussel, R., Czernichow, S., et al. (2012). Rare MTNR1B variants impairing melatonin receptor 1B function contribute to type 2 diabetes. Nat. Genet. *44*, 297–301.

6. Romeo, S., Yin, W., Kozlitina, J., Pennacchio, L.A., Boerwinkle, E., Hobbs, H.H., and Cohen, J.C. (2009). Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. J. Clin. Invest. *119*, 70–79.

7. MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B., et al. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. Science *335*, 823–828.

8. Ahituv, N., Kavaslar, N., Schackwitz, W., Ustaszewska, A., Martin, J., Hebert, S., Doelle, H., Ersoy, B., Kryukov, G., Schmidt, S., et al. (2007). Medical sequencing at the extremes of human body mass. Am. J. Hum. Genet. *80*, 779–791.

9. Ji, W., Foo, J.N., O'Roak, B.J., Zhao, H., Larson, M.G., Simon, D.B., Newton-Cheh, C., State, M.W., Levy, D., and Lifton, R.P. (2008). Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nat. Genet. *40*, 592–599.

10. Sunyaev, S.R. (2012). Inferring causality and functional significance of human coding DNA variants. Hum. Mol. Genet. *21*, R10–7.

11. Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

12. National Heart, Lung, and B.I. Exome Variant Server.

13. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.

14. Ng, P.C., and Henikoff, S. (2001). Predicting deleterious amino acid substitutions. Genome Res. *11*, 863–874.

15. Schwarz, J.M., Rödelsperger, C., Schuelke, M., and Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. Nat. Methods *7*, 575–576.

16. Kiezun, A., Garimella, K., Do, R., Stitziel, N.O., Neale, B.M., McLaren, P.J., Gupta, N., Sklar, P., Sullivan, P.F., Moran, J.L., et al. (2012). Exome sequencing and the genetic basis of complex traits. Nat. Genet. *44*, 623–630.

17. Stoddart, D., Heron, A.J., Mikhailova, E., Maglia, G., and Bayley, H. (2009). Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. Proc. Natl. Acad. Sci. U. S. A. *106*, 7702–7707.

18. Korlach, J., Marks, P.J., Cicero, R.L., Gray, J.J., Murphy, D.L., Roitman, D.B., Pham, T.T., Otto, G.A., Foquet, M., and Turner, S.W. (2008). Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. Proc. Natl. Acad. Sci. U. S. A. *105*, 1176–1181.

19. Dela Torre, R., Larkin, J., Singer, A., and Meller, A. (2012). Fabrication and characterization of solid-state nanopore arrays for high-throughput DNA sequencing. Nanotechnology *23*, 385308.

20. Bell, D.C., Thomas, W.K., Murtagh, K.M., Dionne, C.A., Graham, A.C., Anderson, J.E., and Glover, W.R. (2012). DNA base identification by electron microscopy. Microsc. Microanal. *18*, 1049–1053.

21. Do, R., Kathiresan, S., and Abecasis, G.R. (2012). Exome sequencing and complex disease: practical aspects of rare variant association studies. Hum. Mol. Genet. *21*, R1–9.

22. Holm, H., Gudbjartsson, D.F., Sulem, P., Masson, G., Helgadottir, H.T., Zanon, C., Magnusson, O.T., Helgason, A., Saemundsdottir, J., Gylfason, A., et al. (2011). A rare variant in MYH6 is associated with high risk of sick sinus syndrome. Nat. Genet. *43*, 316–320.

23. Jonsson, T., Atwal, J.K., Steinberg, S., Snaedal, J., Jonsson, P. V, Bjornsson, S., Stefansson, H., Sulem, P., Gudbjartsson, D., Maloney, J., et al. (2012). A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. Nature *488*, 96–99.

24. Skol, A.D., Scott, L.J., Abecasis, G.R., and Boehnke, M. (2007). Optimal designs for two-stage genome-wide association studies. Genet. Epidemiol. *31*, 776–788.

25. Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E.H., Kumar, A., Howard, E., Shendure, J., and Turner, D.J. (2010). Target-enrichment strategies for next-generation sequencing. Nat. Methods *7*, 111–118.

26. Liu, L., Sabo, A., Neale, B.M., Nagaswamy, U., Stevens, C., Lim, E., Bodea, C.A., Muzny, D., Reid, J.G., Banks, E., et al. (2013). Analysis of rare, exonic variation amongst subjects with autism spectrum disorders and population controls. PLoS Genet. *9*, e1003443.

27. Albrechtsen, A., Grarup, N., Li, Y., Sparsø, T., Tian, G., Cao, H., Jiang, T., Kim, S.Y., Korneliussen, T., Li, Q., et al. (2013). Exome sequencing-driven discovery of coding polymorphisms associated with common metabolic phenotypes. Diabetologia *56*, 298–310.

28. Purcell, S.M., Moran, J.L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., O'Dushlaine, C., Chambert, K., Bergen, S.E., Kähler, A., et al. (2014). A polygenic burden of rare disruptive mutations in schizophrenia. Nature *506*, 185–190.

29. Cruchaga, C., Karch, C.M., Jin, S.C., Benitez, B.A., Cai, Y., Guerreiro, R., Harari, O., Norton, J., Budde, J., Bertelsen, S., et al. (2014). Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. Nature *505*, 550–554.

30. Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., and Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748–752.

31. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.

32. Chan, Y., Lim, E.T., Sandholm, N., Wang, S.R., McKnight, A.J., Ripke, S., Daly, M.J., Neale, B.M., Salem, R.M., and Hirschhorn, J.N. (2014). An Excess of Risk-Increasing Low-Frequency Variants Can Be a Signal of Polygenic Inheritance in Complex Diseases. Am. J. Hum. Genet. *94*, 437–452.

33. Agarwala, V., Flannick, J., Sunyaev, S., and Altshuler, D. (2013). Evaluating empirical bounds on complex disease genetic architecture. Nat. Genet.

Appendix

Supplementary Material – Large-scale pooled next-generation sequencing of 1077 genes to identify genetic causes of short stature

Supplemental Methods – Variant Calling Strategy

Variant calling was performed using Syzygy software and then applied a new likelihood-based secondary calling strategy that integrated the extra information from our matrix design. Instead of setting a hard cutoff for calling variants in individual pools as typically implemented in Syzygy, we calculated accumulated evidence for the presence of the variant allele from the full matrix. To keep the false positive rate low, we also required supportive evidence for an individual variant to be present in both row and column pools.

Specifically, our single nucleotide polymorphism (SNP) calling strategy is a likelihood ratio test comparing the following two hypotheses: 1. Null Hypothesis (H_0) – there is no variant at the site; observed non-reference reads are all due to sequencing error. 2. Alternative Hypothesis (H_1) – there is a variant at the site (as well as background sequencing error rate). First, a sequencing error rate is chosen to optimize the likelihood of the null hypothesis ($L[H_0]$); then a background sequencing error rate as well as variant frequencies (in pools where there are reads of the variant allele) are chosen to optimize the likelihood of the alternative hypothesis ($L[H_1]$). The test statistic D is calculated as: $D = 2*(\ln(L[H_1]) - \ln(L[H_0]))$. The probability distribution of D is approximately a chi-square distribution with degrees of freedom equal to $df[H_1]-df[H_0]$ (in our case, the degrees of freedom is equal to the number of pools with non-zero estimated variant frequency). We can then derive the p-values and significance was set at 0.05 after a strict Bonferroni correction accounting for the number of sites tested.

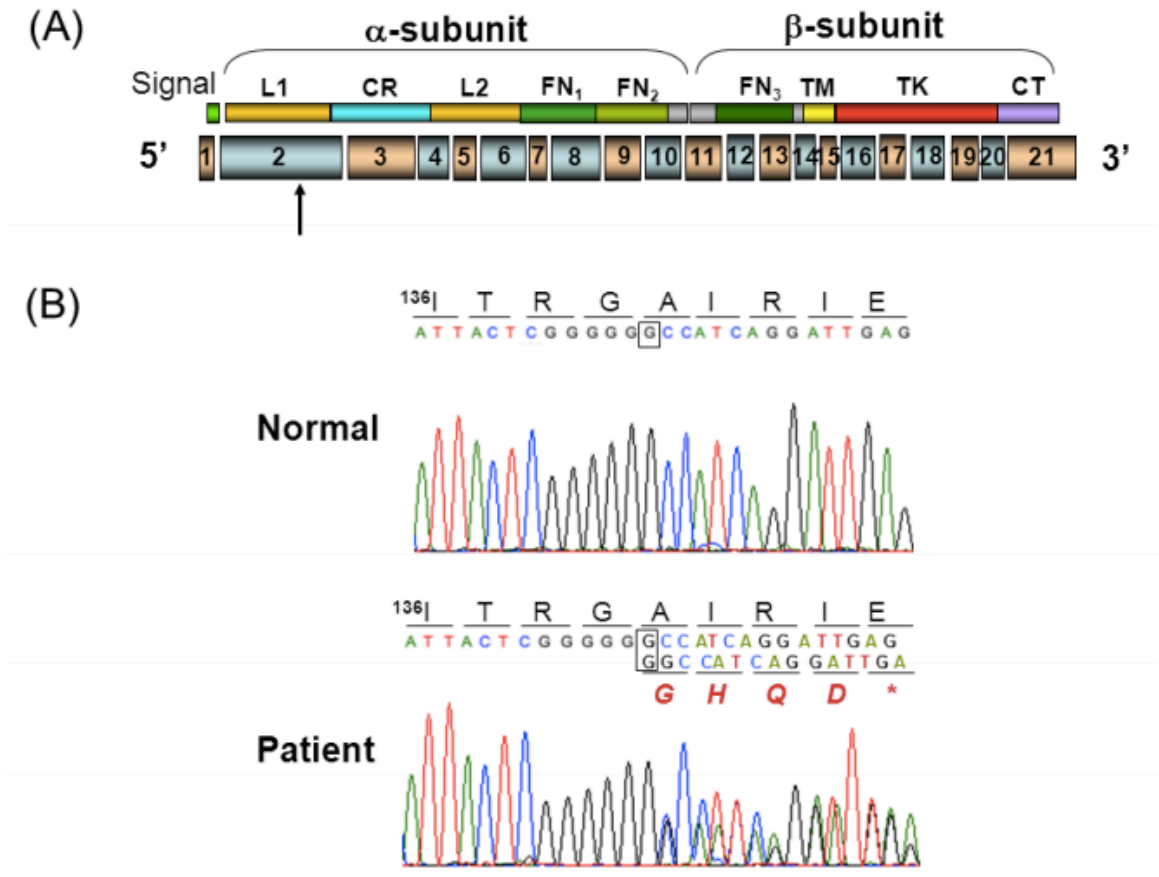

Figure S2.1: Electropherogram of novel IGF1R variant #1, c.418dupG (see Table 2.3). (A) Schematic of the IGF1R peptide is shown above the correlating coding exons (numbered 1 to 21). Arrow indicates the location of the variant. The IGF1R peptide consist of a signal peptide, alpha and beta subunits, composed of the following subdomains: L1, receptor leucine-rich domain; CR, cysteine-rich, furin-like, domain; L2, receptor L domain; FN, fibronectin type III domain; TM, transmembrane domain; TK, tyrosine kinase catalytic domain; and CT, C-terminal domain. (B) Electropherogram of IGF1R c.418dupG (exon 2) in Patient genomic DNA. Box nucleotide, c.418. Normal and new residues generated by the c.418dupG, are as indicated. Asterix (*), stop codon.

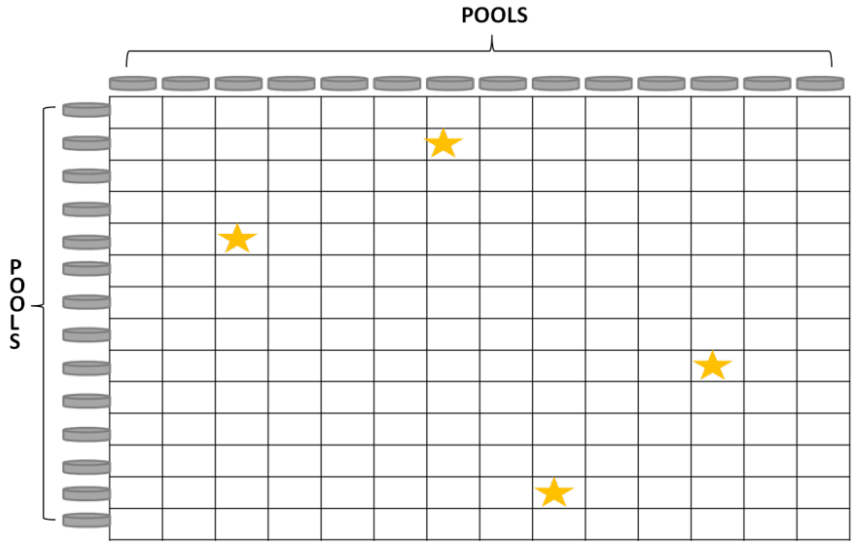

Figure S2.2: Matrix pooling design. Each box in the table represents one sample, which is present in one column pool and one row pool. Four empty "holes", labeled as yellow stars, were included into each matrix for assessing false positive rate.

Table S2.1: Human Gene Mutation Database Dominant Disease Associated Variants Found in Cases Only

Gene	Disease	HGMD Variant ID	Genomic Location (hg19)	c.DNA	Protein	Ref.	Subject Description
PTPN11	Noonan syndrome	CM024733	chr12:112910844	NM_002834.3: c.853T-C	NP_002825.3: p.F285L	[1]	See main text.
PTPN11	Noonan syndrome	CM021136	chr12:112915526	NM_002834.3: c.925A-G	NP_002825.3: p.I309V	[1]	See main text.
PTPN11	Noonan syndrome	CM013416	chr12:112888172	NM_002834.3: c.188A-G	NP_002825.3: p.Y63C	[2]	See main text.
TRPV4	Brachyolmia	CM083203	chr12:110230201	NM_021625.4: c.1858G-A	NP_067638.3: p.V620I	[3]	See main text.
NLRP3	Familial cold auto-inflammatory syndrome	CM043481	chr1:247588214	NM_004895.4: c.1469G-A	NP_004886.3: p.R490K	[4]	ISS. No report of recurrent rash or arthralgias with cold exposure.
CREBBP	Rubinstein-Taybi syndrome	CM021087	chr16:3778387	NM_004380.2: c.6661A-C	NP_004371.2: p.M2221L	[5]	GHD. Normal intelligence. No physical signs of Rubinstein-Taybi.
CREBBP	Rubinstein-Taybi syndrome	CM085347	chr16:3820773	NM_004380.2: c.2678C-T	NP_004371.2: p.S893L	[6]	GHD. Normal intelligence. No physical signs of Rubinstein-Taybi.
GLI3	Greig cephalopoly-syndactyly syndrome	CM970684	chr7:42007506	NM_000168.5: c.2119C-T	NP_000159.3: p.P707S	[7]	ISS, developmental delay. Normal head shape and no polydactyly or syndactyly.

Table S2.1 (Continued)

GLI3	Greig cephalopoly-syndactyly syndrome	CM990707	chr7:42007201	NM_000168.5: c.2424A-G	NP_000159.3: p.I808M	[8]	Dysplastic thumb with mild short stature. Variant does not segregate with phenotype in the family.
FGFR1	Non-syndromic trigono-cephaly	CM010303	chr8:38282064	NM_023110.2: c.899T-C	NP_075598.2: p.I300T	[9]	ISS. Normocephalic.
PTCH1	Holoprosence phaly	CM020752	chr9:98229479	NM_000264.3: c.2479A-G	NP_000255.2: p.S827G	[10]	Found in the subject with the novel *IGF1R* frameshift. No midline defects although has a bilateral cleft lip and palate. No GHD.
JAG1	Alagille syndrome	CM061804	chr20:10622447	NM_000214.2: c.2666G-A	NP_000205.1: p.R889Q	[11]	Patient has clinical diagnosis of Rubinstein-Taybi syndrome. No hepatic abnormalities.
RPL5	Diamond-Blackfan anemia	CM086904	chr1:93301840	NM_000969.3: c.418G-A	NP_000960.2: p.G140S	[12]	ISS. No cleft palate, thumb abnormalities, or anemia.
COL1A1	Osteogenesis imperfecta I	CM123299	chr17:48264220	NM_000088.3: c.3595A-G	NP_000079.2: p.S1199G	[13]	ISS. No history of fractures.
COL1A1	Ehlers-Danlos Syndrome	CM071624	chr17:48265329	NM_000088.3: c.3277C-T	NP_000079.2: p.R1093C	[14]	ISS. No skin hyperextensibility or joint hypermobility.
EXT1	Multiple osteo-chondromas	CM099178	chr8:118830697	NM_000127.2: c.1609G-A	NP_000118.2: p.V537I	[15]	ISS. No osteochondromas.
INSR	Diabetes	CM950700	chr19:7141798	NM_000208.2: c.2572A-G	NP_000199.2: p.T858A	[16]	SGA. No signs of diabetes or insulin resistance.

Table S2.1 (Continued)

COL11A1	Robin Sequence and Marshall syndrome	CS030538	chr1:103400613	NM_080629.2: c.IVS45+3G4A	Unknown. Predicted splice variant.	[17]	Clinically diagnosed with Coffin-Lowry syndrome. No cleft palate.
ALPL	Hypo-phosphatasia	CM084852	chr1:21890638	NM_000478.4: c.577C-G	NP_000469.3: p.P193A	NP	ISS. No bone abnormalities. Normal alkaline phosphatase levels.
ALPL	Hypo-phosphatasia	CM993530	chr1:21890587	NM_000478.4: c.526G-A	NP_000469.3: p.A176T	[18]	ISS. No bone abnormalities. Normal alkaline phosphatase levels.
ABCC8	Hyper-insulinism	CM994416	chr11:17450177	NM_000352.3: c.1858C-T	NP_000343.2: p.R620C	[19]	Hypopituitarism. No signs of hyperinsulinism.
PHEX	Hypo-phosphatemic Rickets	CM025296	chrX:22051133	NM_000444.4: c.10G-C	NP_000435.3: p.E4Q	NP	ISS. Normal phosphorus levels and no rickets.
NF1	Neurofibromatosis 1	CM000785	chr17:29552261	NM_000267.3: c.1994C-T	NP_000258.1: p.S665F	[20]	GHD. No stigmata of neurofibromatosis.
NF1	Neurofibromatosis 1	CS040852	chr17:29653237	NM_000267.3: c.5172G>A	Unknown. Predicted splice variant.	[21]	
BMP4	Orbicularis oris defect	CM091521	chr14:54417117	NM_001202.3: c.860G-A	NP_001193.2: p.R287H	[22]	
TGFBR2	Marfan Syndrome	CM063202	chr3:30733044	NM_003242.5: c.1657T-A	NP_003233.4: p.S553T	[23]	
LRP5	Osteoporosis, primary	CM122876	chr11:68193464	NM_002335.2: c.3446T-A	NP_002326.2: p.L1149Q	[24]	

ISS – Idiopathic short stature. GHD – Growth hormone deficiency. SGA – Small for gestational age. NP – Not published.

Table S2.2: List of 1077 height candidate genes, including (1) genes known to underlie syndromic growth disorders or skeletal dysplasias, as well as genes involved in growth plate biology or growth hormone signaling (bold) (2) genes within genomic loci associated with height based on genome-wide association studies (plain) (3) genes belonging to both group (1) and (2) (underlined).

AAMP	CDSN	GABRD	LIN28	PHOSPHO1	SMARCA5
ABCB5	CENPO	GALNS	LIN28B	PIGC	**SMARCAL1**
ABCB8	CENPP	GAP43	**LMBR1**	PIGF	**SMC1A**
ABCC8	CEP120	GAS1	**LMNA**	PIGK	**SMC3**
ABCE1	**CEP290**	GCNT1	LMO4	PITPNM2	**SMO**
ABI3	CETN3	GDF5	**LMX1B**	PITX1	SMOX
ABP1	CHCHD7	GDPD5	LOXL1	**PITX2**	**SMPD1**
ACAN	CHD1L	GFM1	LPAR1	PJA2	SMS
ACBD4	**CHD7**	GFPT2	LPGAT1	PKIA	**SNAP47**
ACPL2	CHMP4A	GGT7	LRIG3	PKN2	SNED1
ACSS2	CHMP4B	GH1	**LRP5**	PLAG1	SNF8
ACTN1	CHRNG	GH2	LRP6	**PLAGL1**	SNRPC
ACVR1	**CHST3**	**GHR**	LRRC37B	PLCD3	SOCS2
ACY3	CISH	**GHRH**	LSAMP	PLD1	**SOCS3**
ADA	**CLCN5**	**GHRHR**	LTBP1	PLEKHA5	SOCS5
ADAM28	CLDN22	**GHRL**	LTBP2	PLEKHJ1	**SOS1**
ADAMTS10	CLIC4	GHSR	LTBP3	**PLOD2**	**SOST**
ADAMTS12	CNOT6	GIP	LTK	PLXNC1	**SOX2**
ADAMTS17	**COL10A1**	GIPC2	LUM	PML	**SOX3**
ADAMTS2	COL11A1	GIT1	LUZP1	PMPCA	**SOX5**
ADAMTS3	**COL11A2**	**GJA1**	LXN	PNKD	**SOX6**
ADAMTSL2	**COL1A1**	GJE1	LY86	PNPT1	SOX9
ADAMTSL3	**COL1A2**	GKAP1	LYPD1	POLR2B	SPAG9
ADCY3	**COL2A1**	**GLB1**	LYPLAL1	POLR3A	SPDEF
ADCY4	**COL5A1**	GLI2	LYSMD3	POLR3G	**SPG20**
ADRBK1	**COL5A2**	**GLI3**	LYSMD4	POMC	SPINK2
AEBP2	**COL9A1**	GLT25D2	MACC1	**POR**	SRRM1
AGPS	**COL9A2**	GMPR2	MAL2	**POU1F1**	**SRY**
AGXT	**COL9A3**	GNA12	**MAP2K1**	POU5F1	SSH3
AK5	**COMP**	**GNAS**	**MAP2K2**	PPA2	SSR1
AKAP7	COPA	**GNPAT**	MAP2K3	PPAP2A	SSSCA1
ALG12	COX18	GNPTAB	MAP6	PPM1A	**SST**
ALMS1	COX7A2	**GNRH1**	**MAPK1**	PPP2R3A	ST3GAL1
ALPL	CPAMD8	**GNRHR**	**MAPK3**	PPP2R5A	STAG1
ALPP	CPEB4	GPBAR1	MAPK9	**PQBP1**	STARD3NL
ALPPL2	CPN1	**GPC3**	MAPKAPK3	PRAM1	**STAT3**
ALX4	CRADD	GPC5	**MATN3**	PRB1	**STAT5B**

Table S2.2 (Continued)

AMZ1	CREB5	GPC6	MBOAT1	PRDM12	STAU1
ANKH	**CREBBP**	GPR111	MBTD1	PRDM6	STC1
ANKRD13B	CRIPT	GPR115	MC3R	PREPL	STC2
ANKRD13D	CROCC	GPR126	MC4R	**PRG4**	STK10
ANKRD17	**CRTAP**	GPR135	MCFD2	PRKAB2	STK25
ANKS1A	CSE1L	GPR27	MCM10	PRKCD	STK36
ANO5	CSNK1G3	GPR39	**MECP2**	PRKCZ	STOML1
ANO7	**CTDP1**	GPR98	MED28	PRKG2	SUB1
ANTXR2	CTPS	GPSM1	MEF2A	PROCR	SUCLG2
AP3D1	**CTSK**	GRK7	MEF2C	PROK2	SUZ12
AQP12A	**CUL4B**	GRM4	**MESP2**	**PROKR2**	SV2A
ARFGEF2	**CUL7**	GSDMC	**MEST**	**PROP1**	SYN3
ARHGEF3	CXXC4	GSS	METRN	PSMB3	SYT12
ARL6	**CYP11B1**	GTF2B	METTL13	PSORS1C1	T
ARPC2	CYP19A1	GTF2H5	MFAP2	PTCH1	**TAC3**
ARSB	CYP21A2	**GUSB**	MGAT5	**PTCH2**	**TACR3**
ARSE	CYP27A1	GYPA	**MGP**	**PTEN**	TAF11
ASS1	**CYP27B1**	GYPB	MICA	PTGFR	TAF2
ASTN1	DAAM1	H1FX	**MKKS**	**PTH**	TAL2
ATAD2B	DACT1	H2AFY	MKL2	**PTH1R**	TAOK1
ATG7	DCN	HABP4	**MKS1**	PTHLH	TARS
ATG9B	DDX27	HACE1	MLF1	PTPDC1	TAX1BP1
ATP13A2	DDX4	HAGHL	MLXIP	**PTPN11**	**TAZ**
ATP2B1	DDX6	**HCCS**	**MMP13**	PTPRJ	TBC1D21
ATP5SL	**DHCR24**	HCP5	**MMP2**	PXMP3	**TBCE**
ATP6V0A2	**DHCR7**	HDAC11	MMP24	PXMP4	**TBX1**
ATP7A	DHDDS	HDLBP	**MMP9**	QSOX2	TBX10
ATP8B1	DIS3L2	HEMK1	**MNX1**	**RAB23**	TBX15
ATR	DLEU7	HEPACAM2	MOBKL2A	RAB26	**TBX3**
ATRX	DLG5	**HESX1**	MOS	**RAB3GAP1**	TBX4
ATXN3	**DLL3**	HEXIM1	MPHOSPH9	**RAB3GAP2**	**TBX5**
AURKA	**DLX3**	HEXIM2	MRPL42	RAD23B	TCF19
B3GALTL	DNAJC27	HFE	**MRPS16**	RAD51L1	**TCF4**
B3GNT8	DNM3	HHIP	MRPS6	**RAF1**	**TCOF1**
B4GALT7	DNMBP	**HIF1A**	MSL2	**RAI1**	TEAD1
BAK1	DNMT3A	HINT3	MSLN	RARS	TET2
BBS1	DOCK2	HIST1H2AE	**MSX2**	RASA2	**TGFB1**
BBS10	DOCK3	HIST1H2BG	MTMR11	RASGEF1B	TGFB2
BBS12	DOHH	HIST1H3A	MUSK	RASGRP3	**TGFBR1**
BBS2	DOT1L	HIST1H4A	MUSTN1	RASSF10	**TGFBR2**
BBS4	DPCR1	HIVEP2	MYC	RAX2	**THRA**

Table S2.2 (Continued)

BBS5	DTL	HK3	**MYCN**	RBBP8	**THRB**
BBS7	DYM	HLA-DQA1	MYH7B	RBM15B	TIGD1
BBS9	E2F1	HLA-DQB1	MYO1E	**RBM28**	TIGIT
BCKDHA	E4F1	HLA-DRB1	MYO6	RBM45	TIMP3
BCKDHB	**EBP**	HMGA1	MYO9B	RDH12	TINF2
BCL2L14	ECM2	HMGA2	NARFL	**RECQL4**	TLE3
BCL7A	EDEM2	HNRNPK	**NBN**	REST	TLN2
BCL9	EFEMP1	**HOXA11**	NCK1	RFK	TMBIM1
BICD2	EFHD1	**HOXA13**	NCL	RFT1	TMEM176A
BMP2	**EFNB1**	**HOXD13**	NCOA1	RFWD2	TMEM181
BMP3	EFR3B	**HPRT1**	NCOA6	RFX6	TMEM22
BMP4	EIF2AK3	**HRAS**	NCOR2	RHOD	TMEM30A
BMP5	EIF4E2	HS2ST1	NCSTN	RIPK3	TMEM38B
BMP6	EIF4E3	**HSPG2**	NDUFAF1	RMI1	TMEM91
BMP7	EIF5AL1	HTR1D	NDUFB1	RNF135	TNC
BMPR1A	EIF6	**HYAL1**	NDUFV1	RNF24	**TNFRSF11A**
BMPR1B	ENPP2	**ICK**	NEDD8	RNF7	**TNFRSF11B**
BMPR2	**EP300**	ID4	NEK4	**ROR2**	TNFSF10
BNC2	EPB41L1	**IDUA**	NEU1	RORA	**TNFSF11**
BPIL2	EPB41L2	**IFT80**	NF1	**RPL11**	TNP1
BRAF	EPDR1	**IGBP1**	NFATC4	**RPL35A**	TNPO1
BRCA2	EPHB2	**IGF1**	NFIC	RPL5	TNS1
BRUNOL5	EPRS	IGF1R	**NFKBIA**	RPLP1	TOX
BTK	EPYC	**IGF2**	NHLH1	RPS10	TP53I13
BTN1A1	ERC2	IGF2BP2	**NIPBL**	**RPS17**	TP53I3
BTN2A1	**ERCC2**	IGF2BP3	**NLRP3**	**RPS19**	TP53INP2
BUB1B	**ERCC3**	**IGF2R**	NMB	RPS20	**TP63**
BVES	ERLIN1	**IGFALS**	NMBR	RPS24	TRA2A
C12orf12	**ESCO2**	IGFBP1	NMU	**RPS6KA3**	TRA2B
C12orf65	ESR1	**IGFBP2**	NMUR1	**RPS7**	**TRAPPC2**
C13orf1	**ESR2**	IGFBP3	NOG	RPSAP52	TREH
C14orf149	ETS1	IGFBP4	NOS3	RREB1	TRIM13
C14orf181	ETV6	**IGFBP5**	NOV	RSPO3	**TRIM32**
C14orf39	**EVC**	**IGFBP6**	NPC2	RTF1	**TRIM37**
C16orf79	**EVC2**	IGFBP7	NPFFR2	RUNX2	TRIP11
C18orf45	EXOC1	IHH	NPPC	RUNX3	TRMT11
C19orf36	EXOSC5	**IKBKG**	**NPR2**	RYBP	TRPC4AP
C1orf105	**EXT1**	IL20RB	NPR3	**SALL1**	TRPM5
C1orf21	**EXT2**	IL31	NSD1	**SALL4**	**TRPS1**
C1orf86	FAM101A	IL31RA	**NSDHL**	SAMD3	**TRPV4**
C20orf108	FAM124B	IL7	NSMAF	**SBDS**	TSEN15

Table S2.2 (Continued)

C20orf152	FAM148A	IL8RA	NUCB2	SBNO1	TSSC4
C20orf4	FAM148B	INPP5E	NUDT3	SCAND1	TTC27
C2orf34	FAM164A	**INS**	NUDT8	SCARB1	TTC30A
C2orf44	FAM173A	INSR	NUP160	SCMH1	TTC7A
C2orf54	FAM184B	INTS7	NUSAP1	SCUBE3	TTK
C2orf62	FAM27L	IP6K3	**OBSL1**	SDCCAG3	TTLL4
C2orf79	FAM46A	IPPK	**OCRL**	**SDHA**	TTYH3
C2orf84	FAM49B	IRF1	**OFD1**	SDHB	TULP4
C3orf18	FAM81A	**IRS1**	OIP5	SDR16C5	TWIST1
C3orf31	FAM82A1	**IRS2**	OPN5	SEC11A	TXNDC5
C3orf37	**FANCA**	ISCA2	OPTN	SEC16A	UBE2Z
C3orf63	**FANCB**	ITGB8	OR2K2	**SECISBP2**	**UBR1**
C3orf65	FANCC	ITIH1	OR2Z1	**SEMA3E**	UBXN2A
C4orf14	**FANCD2**	ITIH3	OR4A5	SENP2	UHRF1BP1
C5orf23	FANCE	ITIH4	OR4B1	SENP6	UIMC1
C6orf1	**FANCF**	ITPKA	OR4C12	SERAC1	USE1
C6orf106	**FANCG**	ITPR3	OR4C13	SERPINE2	UTP18
C6orf125	**FANCI**	**JAG1**	OR4C46	SERPINH1	UTP6
C6orf138	**FANCL**	**JAK2**	OSR1	SF3A2	VAMP4
C6orf15	**FANCM**	JAZF1	OSTF1	SF3B4	VANGL2
C6orf173	FAR2	JMJD4	OTUD4	SFMBT1	**VDR**
C6orf191	FARP2	**KAL1**	OTUD7B	SFTA2	**VEGFA**
C7orf11	FBLL1	KBTBD8	**OTX2**	**SH3BP2**	VGLL2
C7orf30	**FBLN1**	KCNE2	PACRGL	SH3GL3	VGLL4
C9orf163	FBLN2	KCNH2	PACSIN1	**SHH**	VPRBP
C9orf40	FBLN5	KCNIP4	PADI2	**SHOX**	VPS13C
C9orf41	**FBN1**	KCNJ1	PANK3	SHOX2	VTA1
C9orf64	**FBN2**	KCNJ11	PAPPA	SHQ1	WDR66
C9orf95	FBXL17	KCNJ12	PAPPA2	**SHROOM4**	WDR73
CA2	FBXO7	KCNJ2	**PAPSS2**	**SIL1**	**WISP3**
CA8	FBXW11	KCNQ1	PAQR5	SIX1	**WNT3**
CABLES1	**FBXW4**	KCNRG	PARN	SIX2	**WNT7A**
CAGE1	FCHO2	**KDM5C**	PARVA	SIX3	WNT9A
CAMK1D	FER	KERA	**PAX3**	SIX6	**WRN**
CARD11	**FGD1**	KHDRBS3	**PAX8**	SKI	WWC2
CASKIN1	**FGF1**	KIAA0317	PCBD2	SLAMF6	YEATS4
CASQ1	**FGF10**	KIAA0368	PCCB	SLC16A7	**ZBTB16**
CATSPER3	FGF18	KIAA0586	PCGF2	SLC22A18	ZBTB20
CBLN3	**FGF2**	**KIAA1279**	**PCNT**	SLC22A4	ZBTB38
CCBL2	**FGF21**	KIF1A	PCSK5	SLC22A5	ZCCHC24
CCDC126	**FGF23**	KIF23	PDE10A	**SLC26A2**	ZCCHC6

Table S2.2 (Continued)

CCDC28B	**FGFR1**	KIF27	PDE11A	**SLC29A3**	ZFAT
CCDC3	**FGFR2**	**KISS1**	PDE3A	**SLC2A2**	ZFP36L1
CCDC49	FGFR3	**KISS1R**	PDIA4	SLC30A10	ZFYVE26
CCDC66	FILIP1	**KRAS**	PDLIM4	**SLC34A3**	**ZMPSTE24**
CCDC78	FKBP1B	L3MBTL3	PEA15	**SLC35C1**	ZNF169
CCDC85A	FKTN	LAMC1	**PEG3**	**SLC35D1**	ZNF341
CCDC91	FLI1	LAMC2	PEX1	SLC35D2	ZNF346
CCDC92	**FLNA**	LARGE	**PEX12**	**SLC37A4**	ZNF366
CCDC99	**FLNB**	LASS3	PEX19	SLC38A9	ZNF367
CCHCR1	FMO5	**LBR**	**PEX3**	SLC39A13	ZNF462
CCNB2	FNDC3B	LCORL	**PEX5**	**SLC4A4**	ZNF652
CCRK	FOLH1	LDHAL6B	**PEX6**	SLC5A3	ZNF664
CD2AP	**FOXC1**	LEMD2	**PEX7**	**SLC6A8**	ZNF678
CD81	FREM3	**LEMD3**	PFN4	SLCO1B3	ZNF683
CDC14B	FRS2	**LEPRE1**	PGP	SLCO1C1	ZNFX1
CDC42EP3	FSD1L	**LFNG**	PHB	SLFNL1	ZSCAN2
CDH3	FUBP3	LHX3	**PHEX**	SLIT2	ZZZ3
CDK2AP1	**FUCA1**	**LHX4**	PHF20	SLIT3	
CDKN1C	**G6PC**	**LIFR**	**PHF6**	SLITRK5	
CDKN2AIP	GAB1	**LIG4**	PHLDB1	SMAD1	

REFERENCES

1. Tartaglia, M., Kalidas, K., Shaw, A., Song, X., Musat, D.L., van der Burgt, I., Brunner, H.G., Bertola, D.R., Crosby, A., Ion, A., et al. (2002). PTPN11 mutations in Noonan syndrome: molecular spectrum, genotype-phenotype correlation, and phenotypic heterogeneity. Am. J. Hum. Genet. *70*, 1555–1563.

2. Tartaglia, M., Mehler, E.L., Goldberg, R., Zampino, G., Brunner, H.G., Kremer, H., van der Burgt, I., Crosby, A.H., Ion, A., Jeffery, S., et al. (2001). Mutations in PTPN11, encoding the protein tyrosine phosphatase SHP-2, cause Noonan syndrome. Nat. Genet. *29*, 465–468.

3. Rock, M.J., Prenen, J., Funari, V.A., Funari, T.L., Merriman, B., Nelson, S.F., Lachman, R.S., Wilcox, W.R., Reyno, S., Quadrelli, R., et al. (2008). Gain-of-function mutations in TRPV4 cause autosomal dominant brachyolmia. Nat. Genet. *40*, 999–1003.

4. Aróstegui, J.I., Aldea, A., Modesto, C., Rua, M.J., Argüelles, F., González-Enseñat, M.A., Ramos, E., Rius, J., Plaza, S., Vives, J., et al. (2004). Clinical and genetic heterogeneity among Spanish patients with recurrent autoinflammatory syndromes associated with the CIAS1/PYPAF1/NALP3 gene. Arthritis Rheum. *50*, 4045–4050.

5. Coupry, I., Roudaut, C., Stef, M., Delrue, M.-A., Marche, M., Burgelin, I., Taine, L., Cruaud, C., Lacombe, D., and Arveiler, B. (2002). Molecular analysis of the CBP gene in 60 patients with Rubinstein-Taybi syndrome. J. Med. Genet. *39*, 415–421.

6. Schorry, E.K., Keddache, M., Lanphear, N., Rubinstein, J.H., Srodulski, S., Fletcher, D., Blough-Pfau, R.I., and Grabowski, G.A. (2008). Genotype-phenotype correlations in Rubinstein-Taybi syndrome. Am. J. Med. Genet. A *146A*, 2512–2519.

7. Wild, A., Kalff-Suske, M., Vortkamp, A., Bornholdt, D., König, R., and Grzeschik, K.H. (1997). Point mutations in human GLI3 cause Greig syndrome. Hum. Mol. Genet. *6*, 1979–1984.

8. Kalff-Suske, M., Wild, A., Topp, J., Wessling, M., Jacobsen, E.M., Bornholdt, D., Engel, H., Heuer, H., Aalfs, C.M., Ausems, M.G., et al. (1999). Point mutations throughout the GLI3 gene cause Greig cephalopolysyndactyly syndrome. Hum. Mol. Genet. *8*, 1769–1777.

9. Kress, W., Petersen, B., Collmann, H., and Grimm, T. (2000). An unusual FGFR1 mutation (fibroblast growth factor receptor 1 mutation) in a girl with non-syndromic trigonocephaly. Cytogenet. Cell Genet. *91*, 138–140.

10. Ming, J.E., Kaupas, M.E., Roessler, E., Brunner, H.G., Golabi, M., Tekin, M., Stratton, R.F., Sujansky, E., Bale, S.J., and Muenke, M. (2002). Mutations in PATCHED-1, the receptor for SONIC HEDGEHOG, are associated with holoprosencephaly. Hum. Genet. *110*, 297–301.

11. Warthen, D.M., Moore, E.C., Kamath, B.M., Morrissette, J.J.D., Sanchez-Lara, P.A., Sanchez, P., Piccoli, D.A., Krantz, I.D., and Spinner, N.B. (2006). Jagged1 (JAG1) mutations in Alagille syndrome: increasing the mutation detection rate. Hum. Mutat. *27*, 436–443.

12. Gazda, H.T., Sheen, M.R., Vlachos, A., Choesmel, V., O'Donohue, M.-F., Schneider, H., Darras, N., Hasman, C., Sieff, C.A., Newburger, P.E., et al. (2008). Ribosomal protein L5 and L11 mutations are associated with cleft palate and abnormal thumbs in Diamond-Blackfan anemia patients. Am. J. Hum. Genet. *83*, 769–780.

13. Zhang, Z.-L., Zhang, H., Ke, Y., Yue, H., Xiao, W.-J., Yu, J.-B., Gu, J.-M., Hu, W.-W., Wang, C., He, J.-W., et al. (2012). The identification of novel mutations in COL1A1, COL1A2, and LEPRE1 genes in Chinese patients with osteogenesis imperfecta. J. Bone Miner. Metab. *30*, 69–77.

14. Malfait, F., Symoens, S., De Backer, J., Hermanns-Lê, T., Sakalihasan, N., Lapière, C.M., Coucke, P., and De Paepe, A. (2007). Three arginine to cysteine substitutions in the pro-alpha (I)-collagen chain cause Ehlers-Danlos syndrome with a propensity to arterial rupture in early adulthood. Hum. Mutat. *28*, 387–395.

15. Jennes, I., Pedrini, E., Zuntini, M., Mordenti, M., Balkassmi, S., Asteggiano, C.G., Casey, B., Bakker, B., Sangiorgi, L., and Wuyts, W. (2009). Multiple osteochondromas: mutation update and description of the multiple osteochondromas mutation database (MOdb). Hum. Mutat. *30*, 1620–1627.

16. Kan, M., Kanai, F., Iida, M., Jinnouchi, H., Todaka, M., Imanaka, T., Ito, K., Nishioka, Y., Ohnishi, T., and Kamohara, S. (1995). Frequency of mutations of insulin receptor gene in Japanese patients with NIDDM. Diabetes *44*, 1081–1086.

17. Melkoniemi, M., Koillinen, H., Männikkö, M., Warman, M.L., Pihlajamaa, T., Kääriäinen, H., Rautio, J., Hukki, J., Stofko, J.A., Cisneros, G.J., et al. (2003). Collagen XI sequence

variations in nonsyndromic cleft palate, Robin sequence and micrognathia. Eur. J. Hum. Genet. *11*, 265–270.

18. Taillandier, A., Cozien, E., Muller, F., Merrien, Y., Bonnin, E., Fribourg, C., Simon-Bouy, B., Serre, J.L., Bieth, E., Brenner, R., et al. (2000). Fifteen new mutations (-195C>T, L-12X, 298-2A>G, T117N, A159T, R229S, 997+2T>A, E274X, A331T, H364R, D389G, 1256delC, R433H, N461I, C472S) in the tissue-nonspecific alkaline phosphatase (TNSALP) gene in patients with hypophosphatasia. Hum. Mutat. *15*, 293.

19. Aguilar-Bryan, L., and Bryan, J. (1999). Molecular biology of adenosine triphosphate-sensitive potassium channels. Endocr. Rev. *20*, 101–135.

20. Fahsold, R., Hoffmeyer, S., Mischung, C., Gille, C., Ehlers, C., Kücükceylan, N., Abdel-Nour, M., Gewies, A., Peters, H., Kaufmann, D., et al. (2000). Minor lesion mutational spectrum of the entire NF1 gene does not explain its high mutability but points to a functional domain upstream of the GAP-related domain. Am. J. Hum. Genet. *66*, 790–818.

21. Mattocks, C., Baralle, D., Tarpey, P., ffrench-Constant, C., Bobrow, M., and Whittaker, J. (2004). Automated comparative sequence analysis identifies mutations in 89% of NF1 patients and confirms a mutation cluster in exons 11-17 distinct from the GAP related domain. J. Med. Genet. *41*, e48.

22. Suzuki, S., Marazita, M.L., Cooper, M.E., Miwa, N., Hing, A., Jugessur, A., Natsume, N., Shimozato, K., Ohbayashi, N., Suzuki, Y., et al. (2009). Mutations in BMP4 are associated with subepithelial, microform, and overt cleft lip. Am. J. Hum. Genet. *84*, 406–411.

23. Mátyás, G., Arnold, E., Carrel, T., Baumgartner, D., Boileau, C., Berger, W., and Steinmann, B. (2006). Identification and in silico analyses of novel TGFBR1 and TGFBR2 mutations in Marfan syndrome-related disorders. Hum. Mutat. *27*, 760–769.

24. Korvala, J., Jüppner, H., Mäkitie, O., Sochett, E., Schnabel, D., Mora, S., Bartels, C.F., Warman, M.L., Deraska, D., Cole, W.G., et al. (2012). Mutations in LRP5 cause primary osteoporosis without features of OI by reducing Wnt signaling activity. BMC Med. Genet. *13*, 26.

Supplementary Material – Simulation of Finnish population history, guided by empirical genetic data, to assess power of rare variant tests in Finland

Supplemental Method

Fitting Finnish demographic history parameters

P(data|model) for each model was calculated as below. The first two terms are the probabilities of the observed allele frequency spectra of synonymous and missense variants given the demographic model being tested. The third term is the probability of the observed synonymous/missense ratio. The last two terms calculate the probabilities of the observed allele sharing between Finns and NFEs.

$$P(data|model) = \binom{s}{s_1, \ \ldots \ , s_6} \prod_{i=1}^{6} {p_i}^{s_i} \cdot \binom{m}{m_1, \ \ldots \ , m_6} \prod_{i=1}^{6} {q_i}^{m_i} \cdot \prod_{i=1}^{6} \binom{s_i + m_i}{s_i} {r_i}^{s_i} (1 - r_i)^{m_i} \cdot \prod_{i=1}^{6} \binom{s_i}{ss_i} {x_i}^{ss_i} (1 - x_i)^{s_i - ss_i} \cdot \prod_{i=1}^{6} \binom{m_i}{sm_i} {y_i}^{sm_i} (1 - y_i)^{m_i - sm_i}$$

s, m: the observed total number of synonymous or missense variants;

s_i, m_i: the observed number of synonymous or missense variants within the i^{th} frequency category;

p_i, q_i: the predicted proportion of synonymous or missense variants that fall into the i^{th} frequency category;

r_i: the predicted proportion of variants in the i^{th} frequency category that are synonymous;

ss_i, sm_i: the observed number of synonymous or missense variants within the i^{th} frequency category in the Finns that are shared with NFEs;

x_i, y_i: the predicted proportion of synonymous or missense variants within the i^{th} frequency category in the Finns that are shared with NFEs.

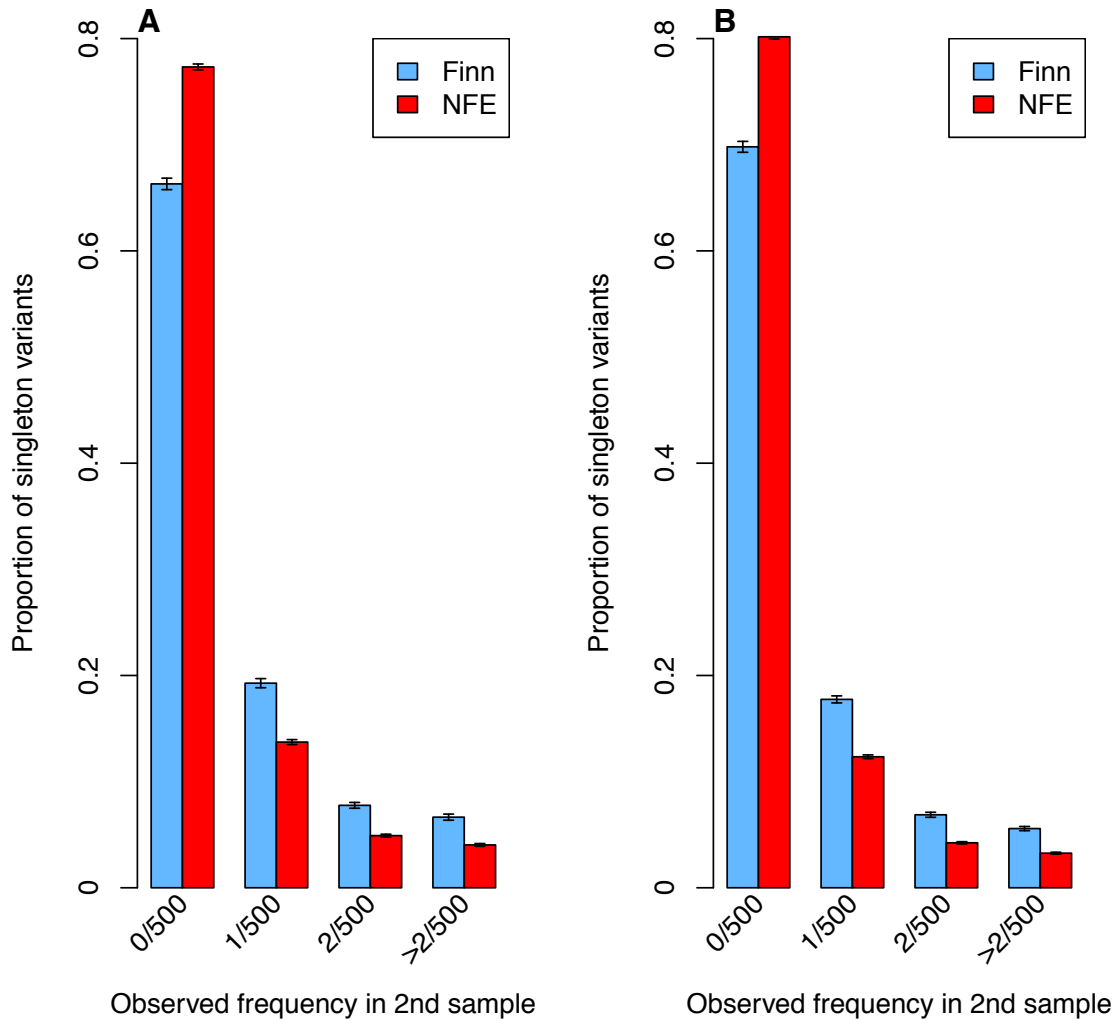

Figure S4.1: Singleton variants in a population of Finns are more likely to be seen again in another population of Finns. For the set of singleton variants ascertained from a random sample of 250 individuals, we assessed the proportion (y-axis) and the frequencies (x-axis) of these variants observed in a second sample of 250 individuals. The analysis was done in synonymous (A) and missense (B) variants separately.

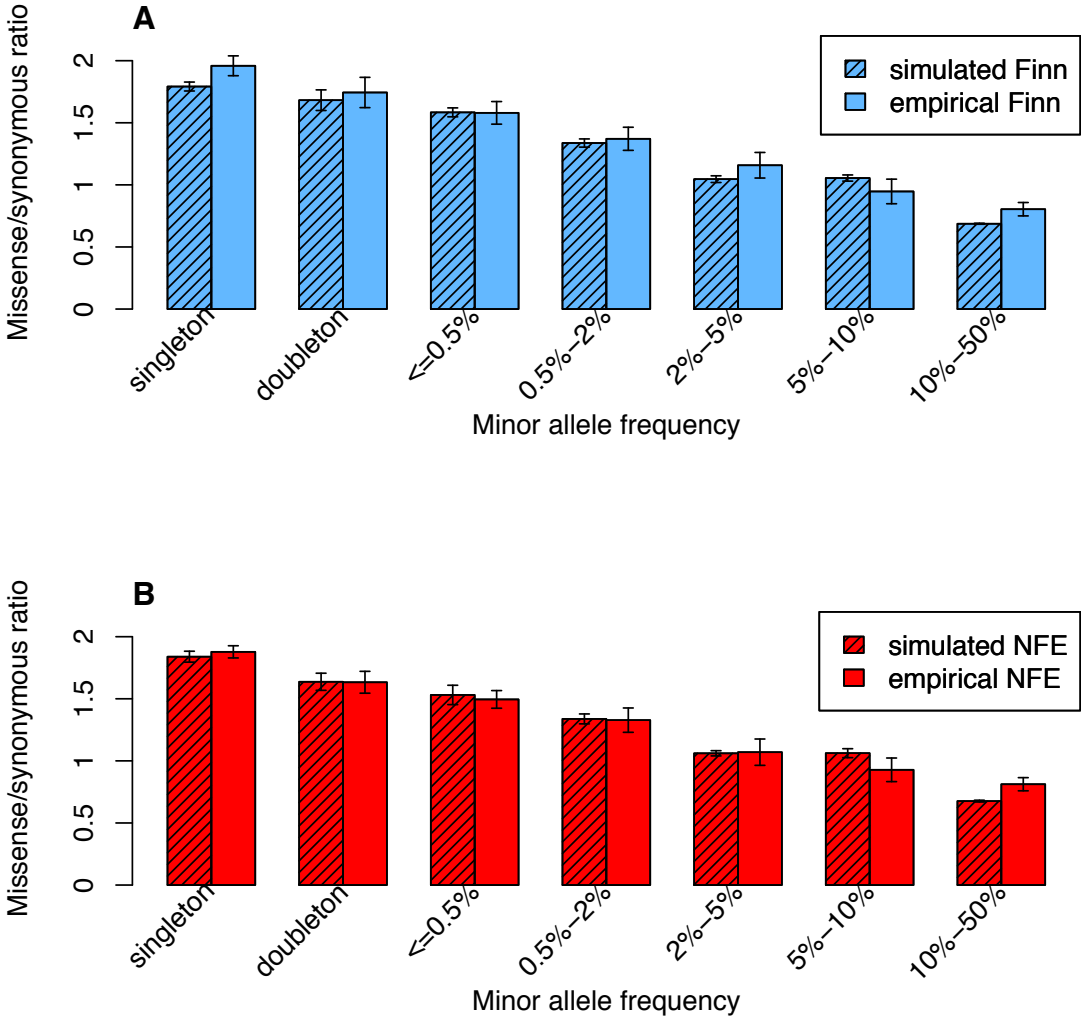

Figure S4.2: Agreement of empirical missense/synonymous ratios with the modeled ratios.

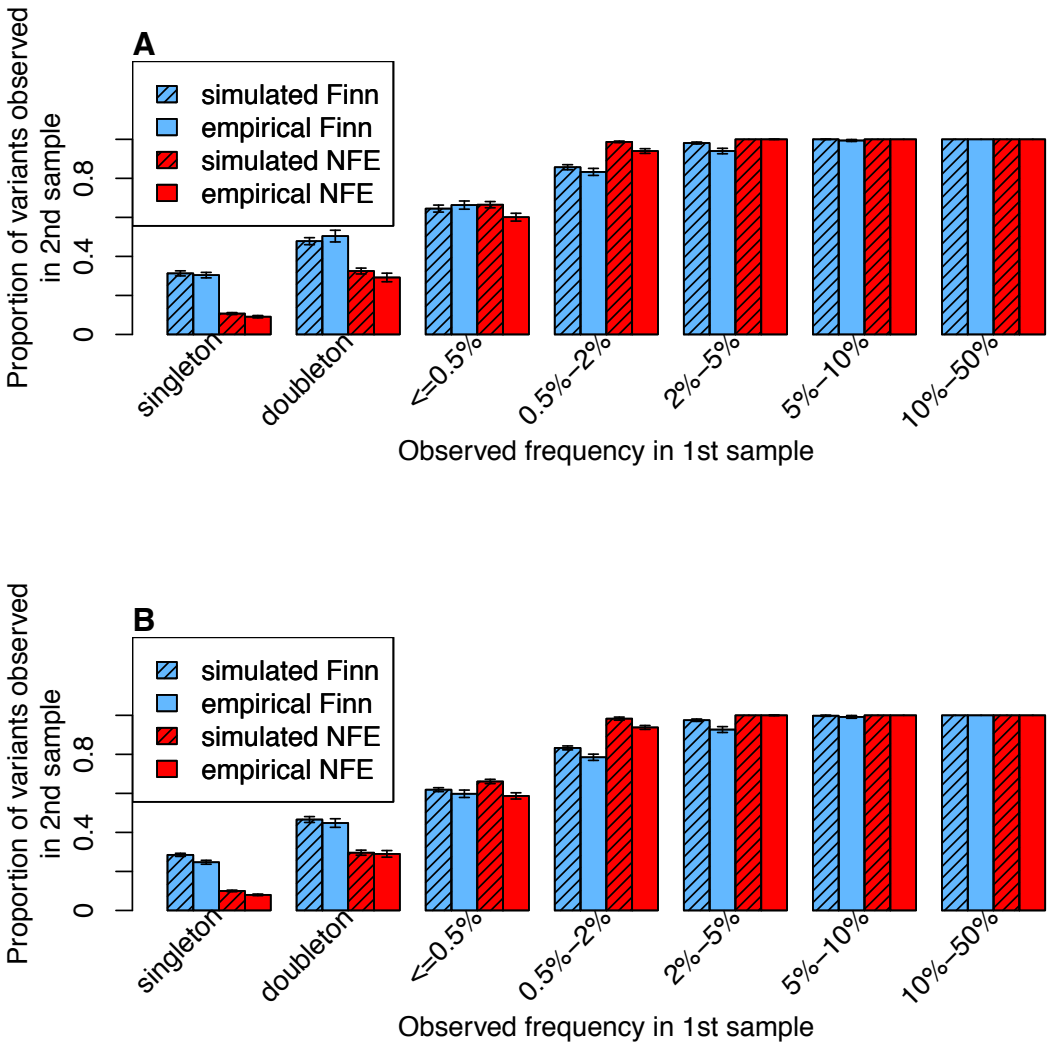

Figure S4.3: Allele sharing between the Finns and the NFEs, comparing simulated data and empirical data. For the set of variants ascertained from the first sample, we assessed their frequencies (x-aixs) in the first sample and the proportion (y-axis) of these variants observed in a second sample. For results in Finns, the first sample is 843 Finns and the second sample is 820 NFEs; for results in NFEs, the first sample is 820 NFEs and the second sample is 843 Finns. The analysis was done in synonymous variants (A) and missense variants (B) separately.

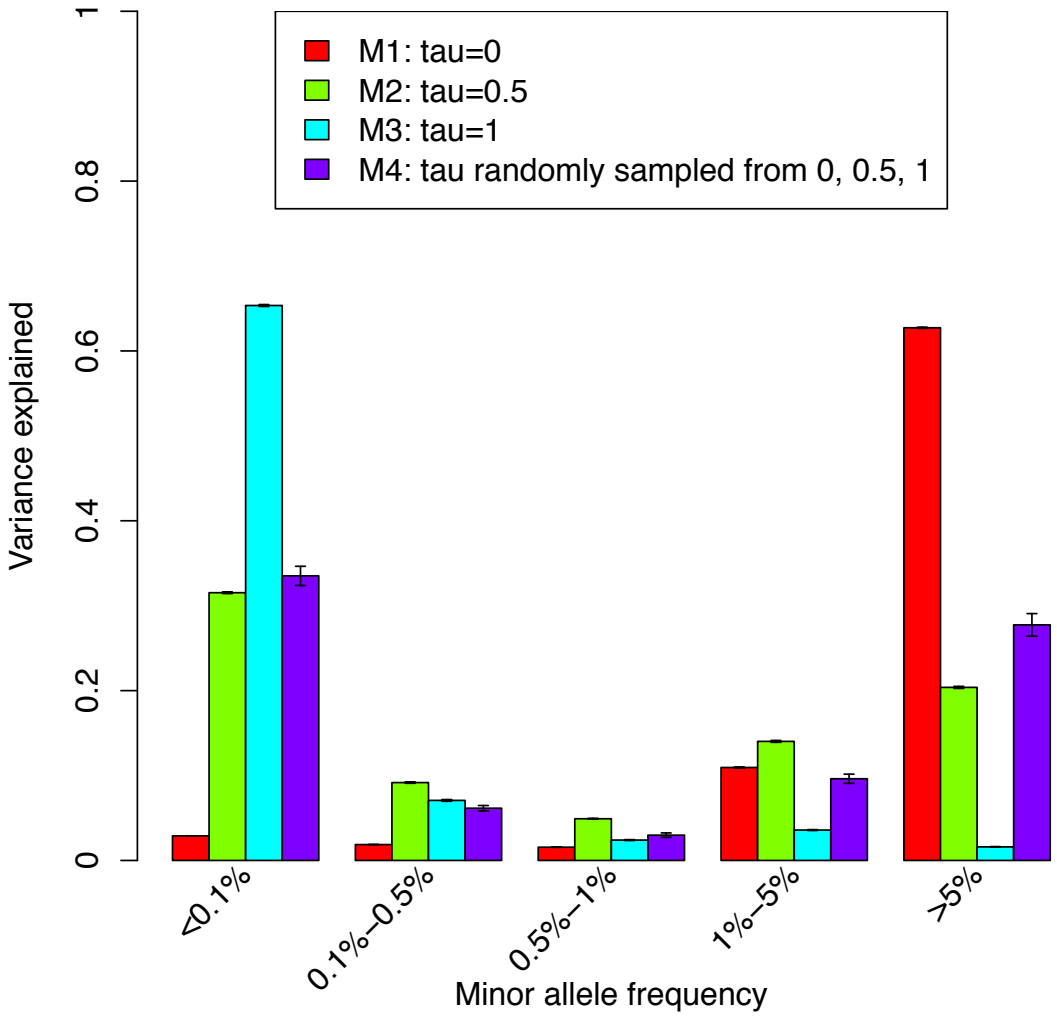

Figure S4.4: Variance explained by variants within different frequency ranges under four different disease models.

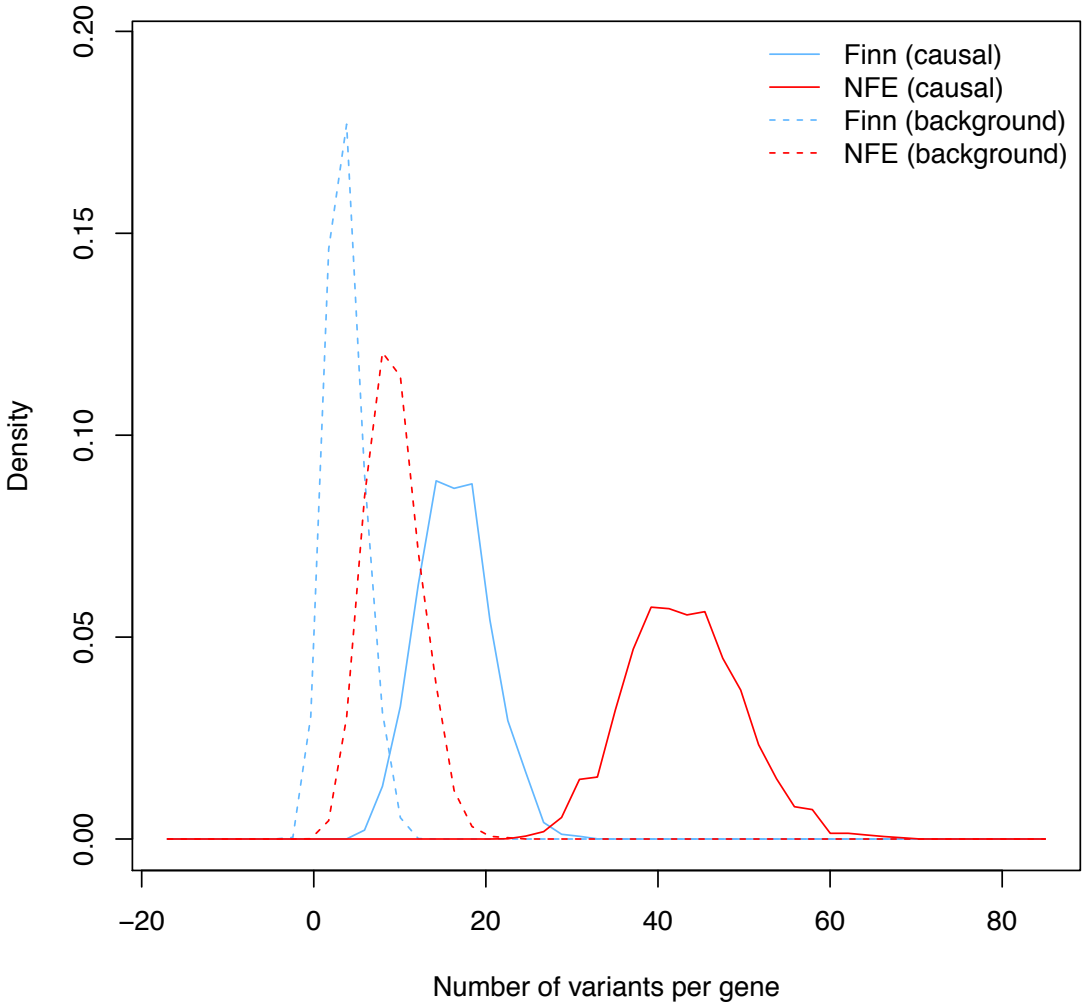

Figure S4.5: Number of causal variants (solid lines) or background variants (dashed lines) with MAF below 5% per gene, in either 30,000 Finns or 30,000 NFEs.

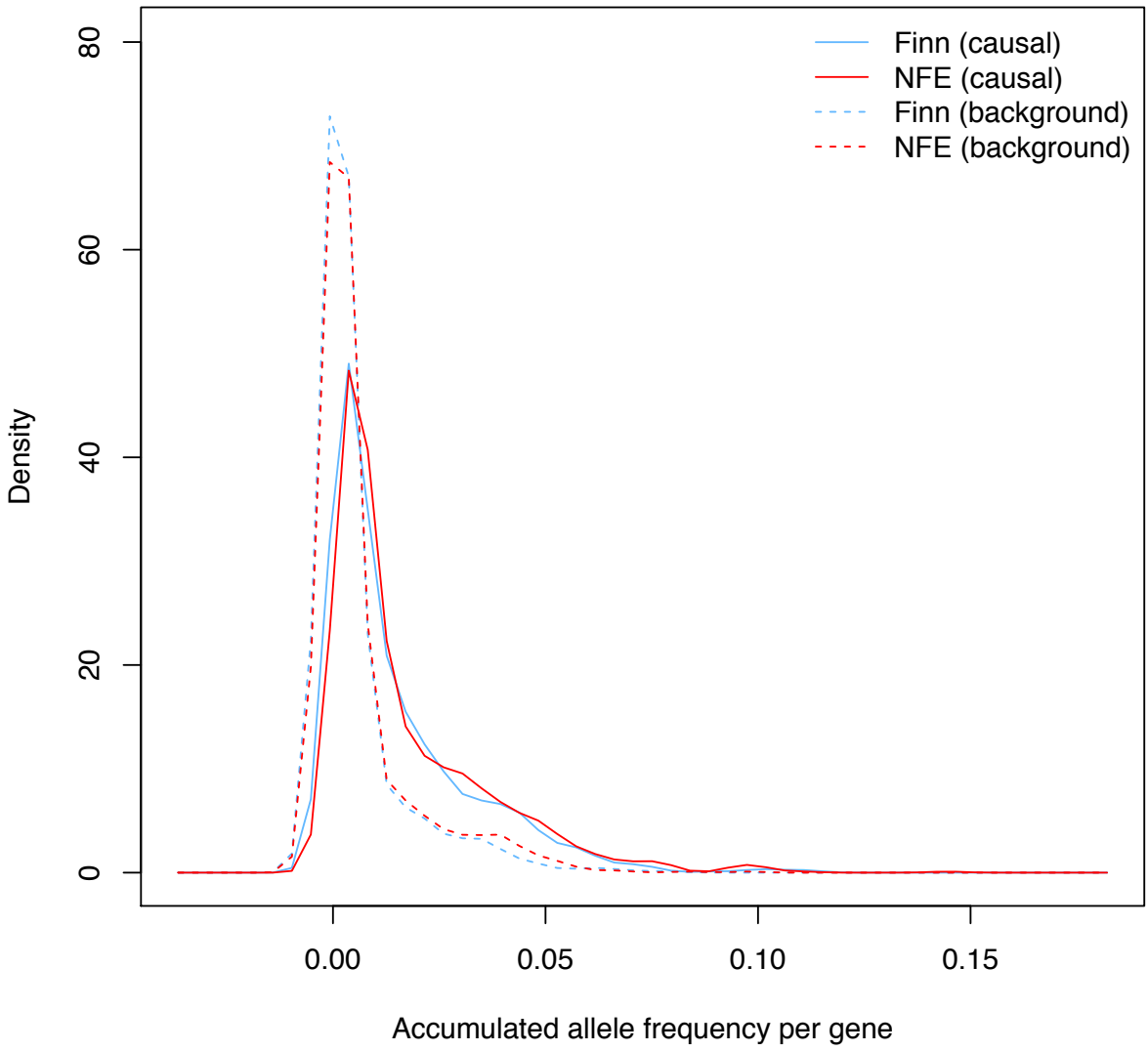

Figure S4.6: Accumulated allele frequency of causal variants (solid lines) or background variants (dashed lines) with MAF below 5% per gene, in either 30,000 Finns or 30,000 NFEs.

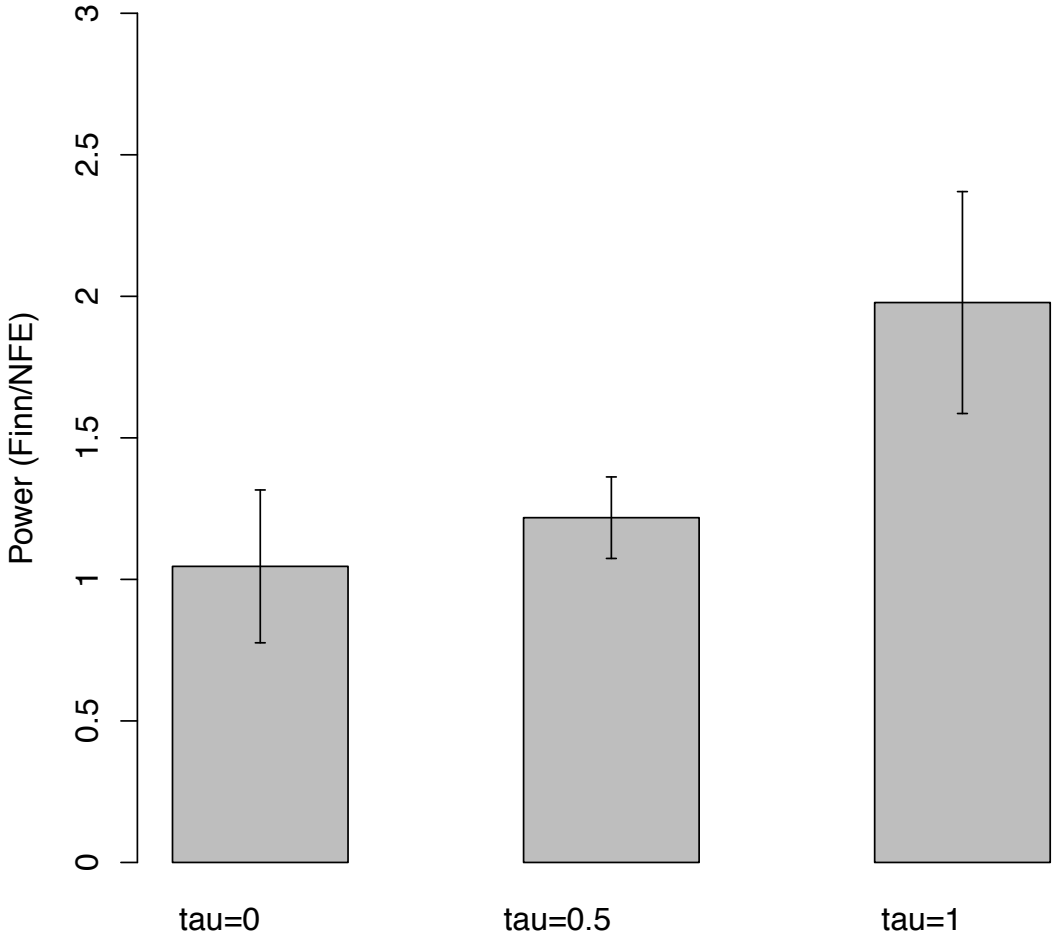

Figure S4.7: Power difference between using the Finns and the NFEs for genes of different τ values under M4 (τ randomly sampled from 0, 0.5, and 1 for each effect gene). Shown here is the result for SKAT-O test and the sample size is 30,000. The biggest power gain in the Finns is seen among genes with τ value of 1 (almost doubling in power, paired t-test p value < 0.01).

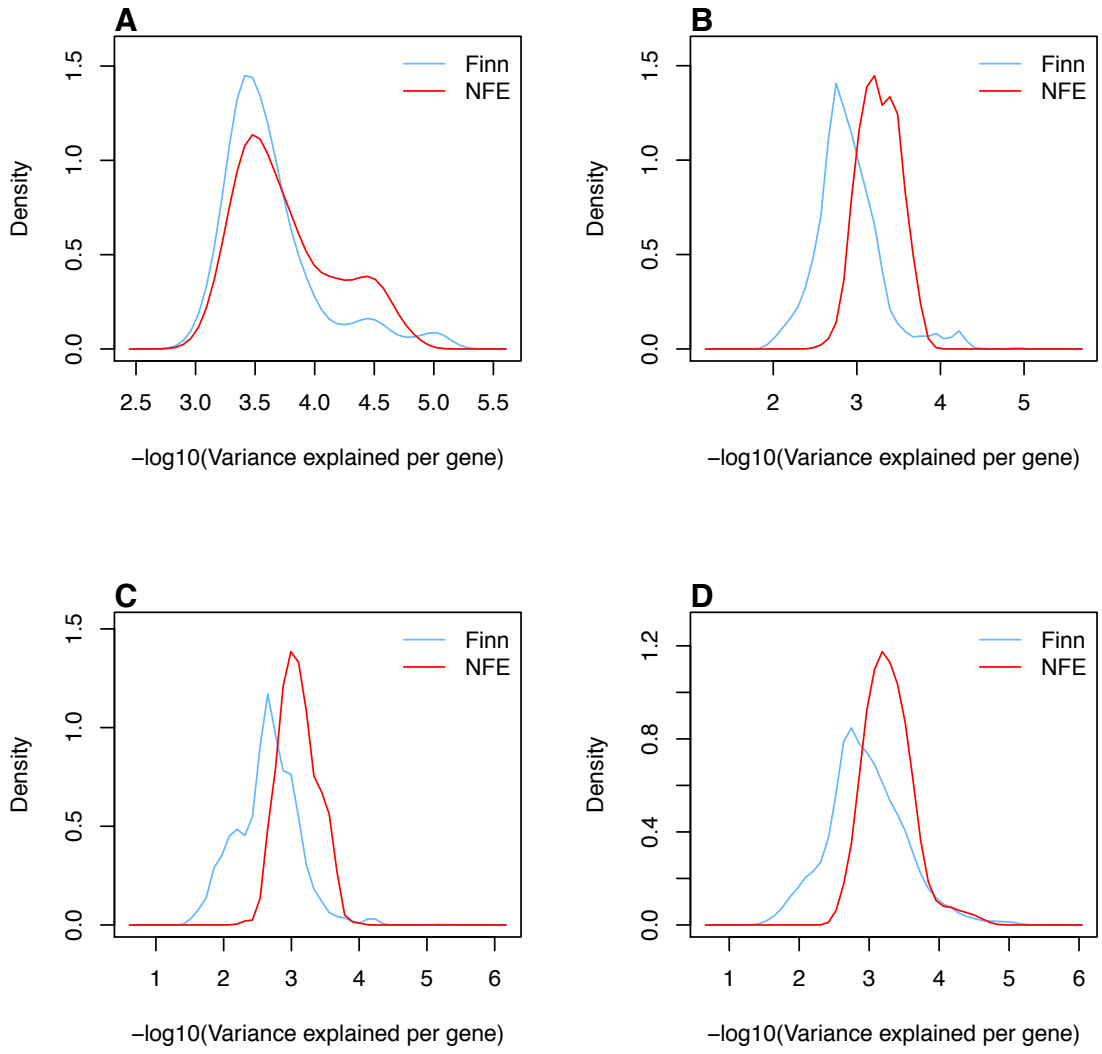

Figure S4.8: Distribution of variance explained per gene by variants with MAF below 5% under four different disease models in either 30,000 Finns or 30,000 NFEs, for genes detected in the Finns only. (A) M1 (τ=0); (B) M2 (τ=0.5); (C) M3 (τ=1); (D) M4 (τ randomly sampled from 0, 0.5, and 1 for each effect gene).

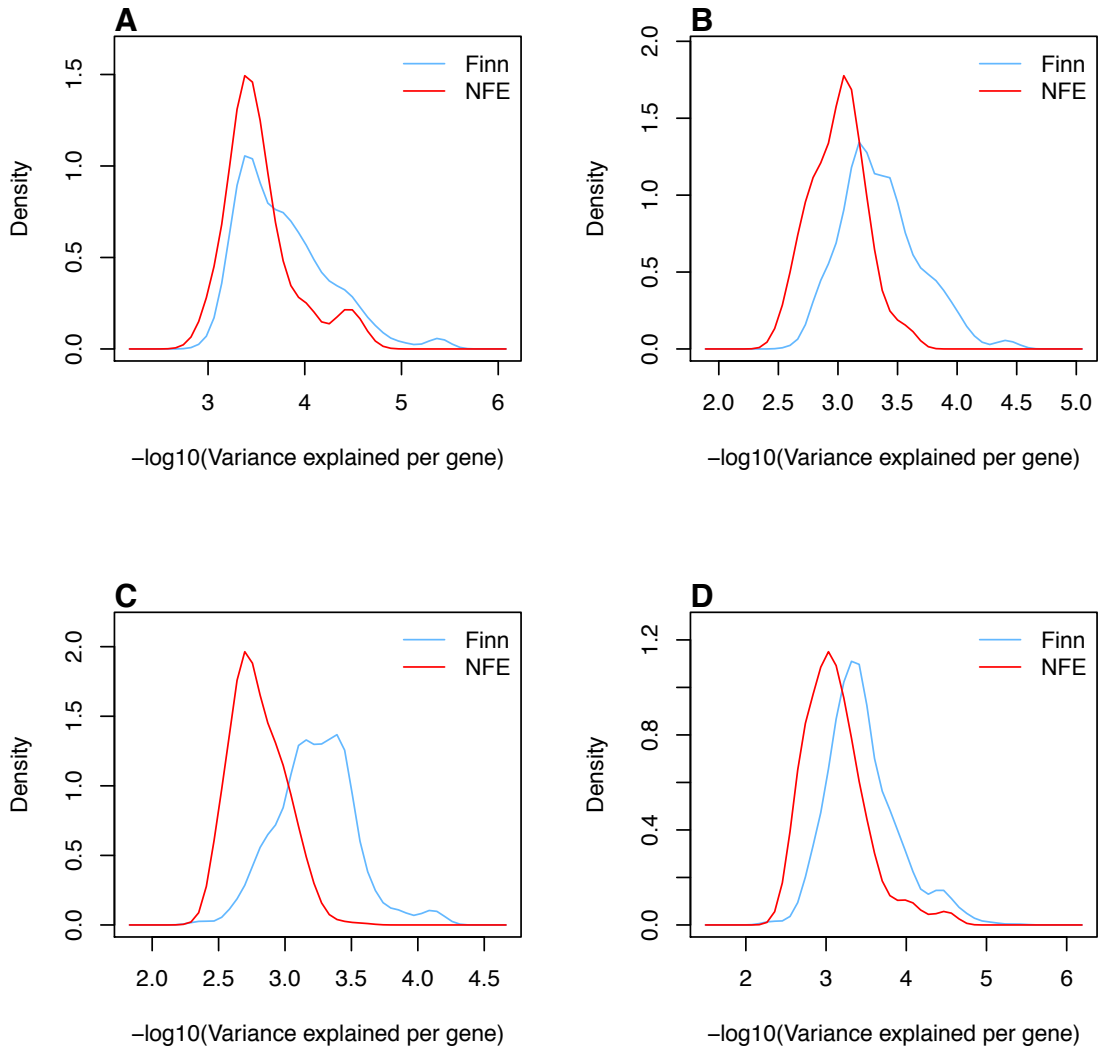

Figure S4.9: Distribution of variance explained per gene by variants with MAF below 5% under four different disease models, in either 30,000 Finns or 30,000 NFEs, for genes detected in the NFEs only. (A) M1 (τ=0); (B) M2 (τ=0.5); (C) M3 (τ=1); (D) M4 (τ randomly sampled from 0, 0.5, and 1 for each effect gene).

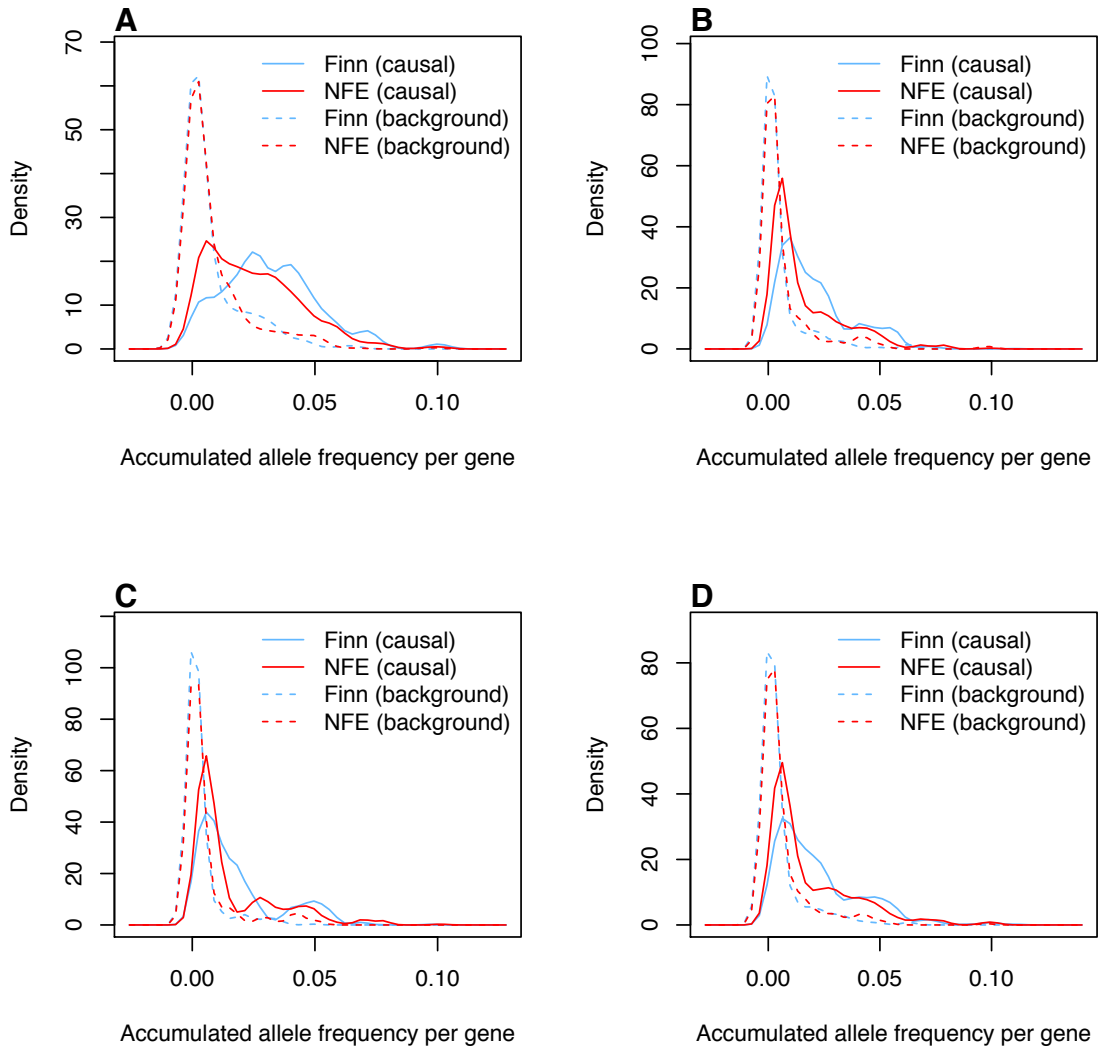

Figure S4.10 Accumulated allele frequency of causal variants (solid lines) or background variants (dashed lines) with MAF below 5% per gene under four different disease models, for genes detected in the Finns only. The distributions for causal variants in the Finns shift upwards compared to the NFEs. (A) M1 (τ=0); (B) M2 (τ=0.5); (C) M3 (τ=1); (D) M4 (τ randomly sampled from 0, 0.5, and 1 for each effect gene).

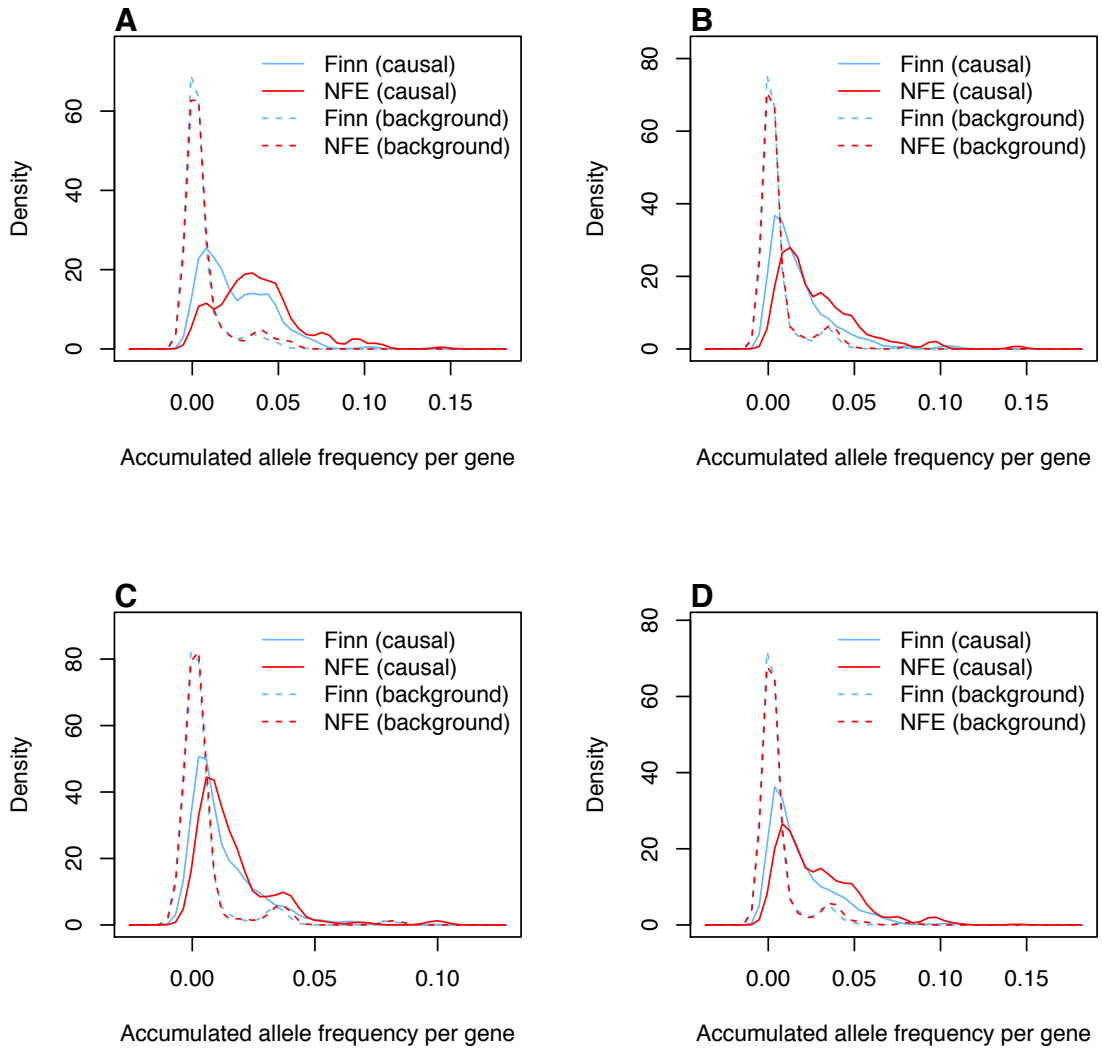

Figure S4.11 Accumulated allele frequency of causal variants (solid lines) or background variants (dashed lines) with MAF below 5% per gene under four different disease models, for genes detected in the NFEs only. The distributions for causal variants in the Finns shift downwards compared to the NFEs. (A) M1 (τ=0); (B) M2 (τ=0.5); (C) M3 (τ=1); (D) M4 (τ randomly sampled from 0, 0.5, and 1 for each effect gene).

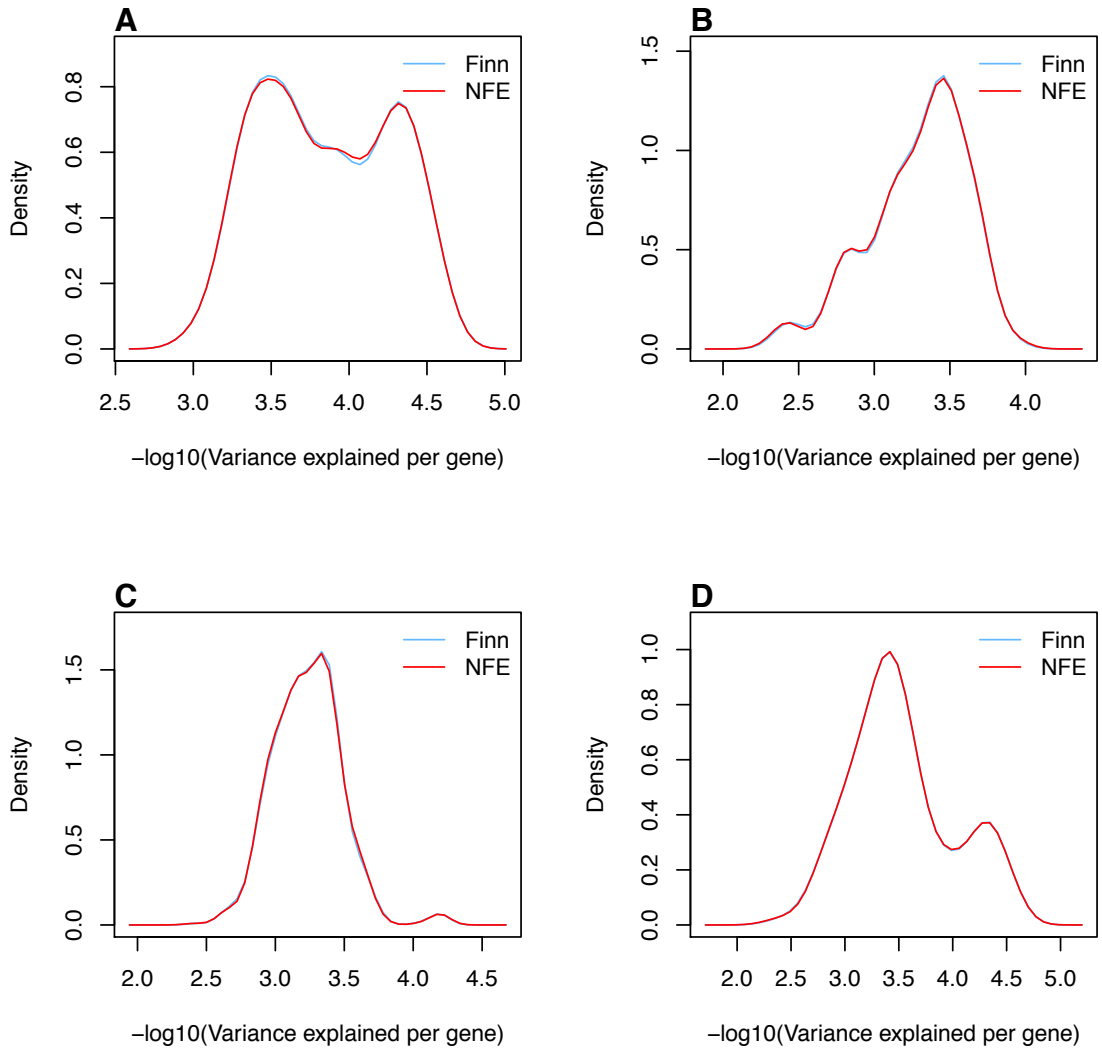

Figure S4.12 Distribution of variance explained per gene by variants with MAF below 5% under four different disease models, in either 30,000 Finns or 30,000 NFEs. The genes were sampled so as to match the variance explained in the Finns and the NFEs. (A) M1 (τ=0); (B) M2 (τ=0.5); (C) M3 (τ=1); (D) M4 (τ randomly sampled from 0, 0.5, and 1 for each effect gene).

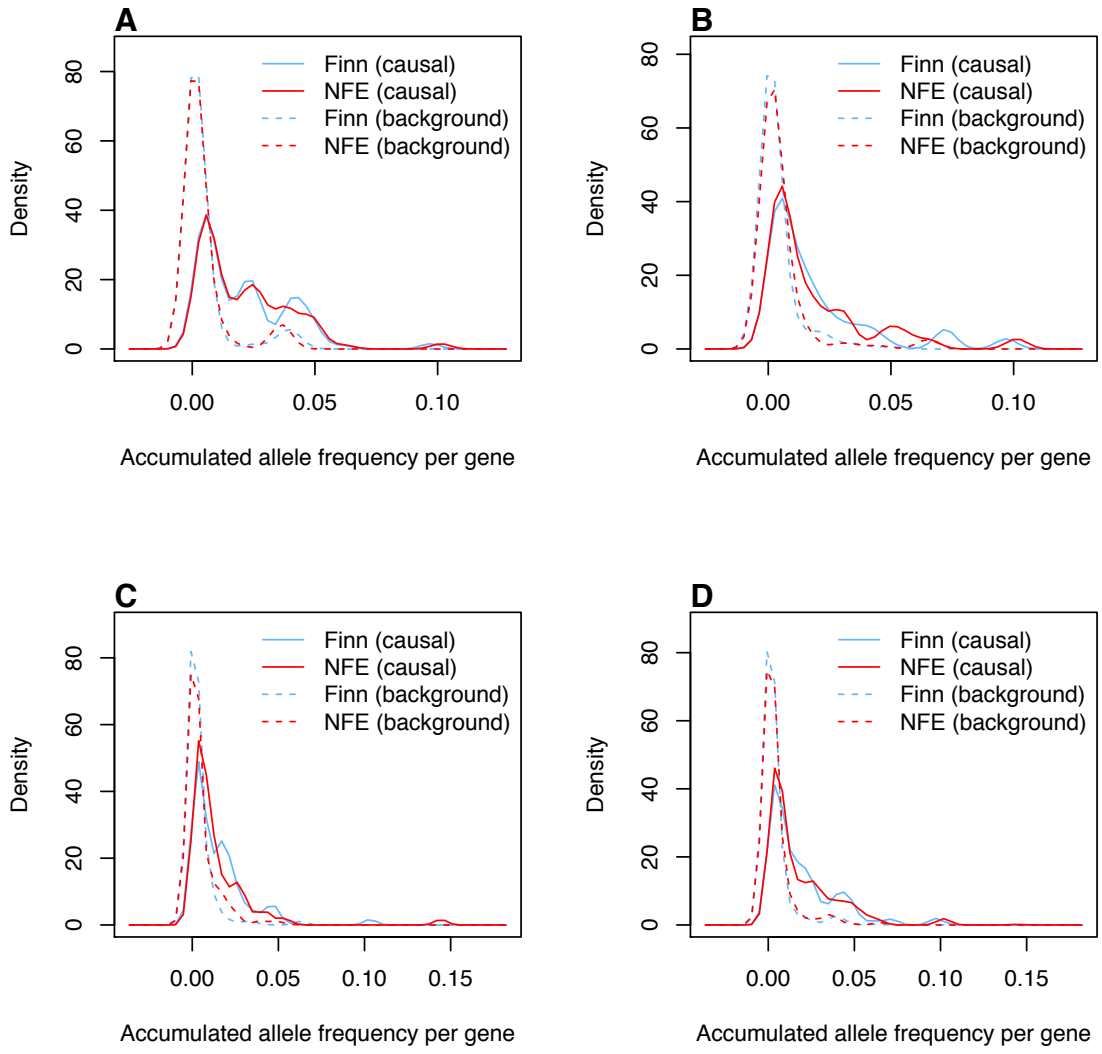

Figure S4.13 Accumulated allele frequency of causal variants (solid lines) or background variants (dashed lines) with MAF below 5% per gene under four different disease models. The genes were sampled so as to match the variance explained in the Finns and the NFEs. (A) M1 (τ=0); (B) M2 (τ=0.5); (C) M3 (τ=1); (D) M4 (τ randomly sampled from 0, 0.5, and 1 for each effect gene).

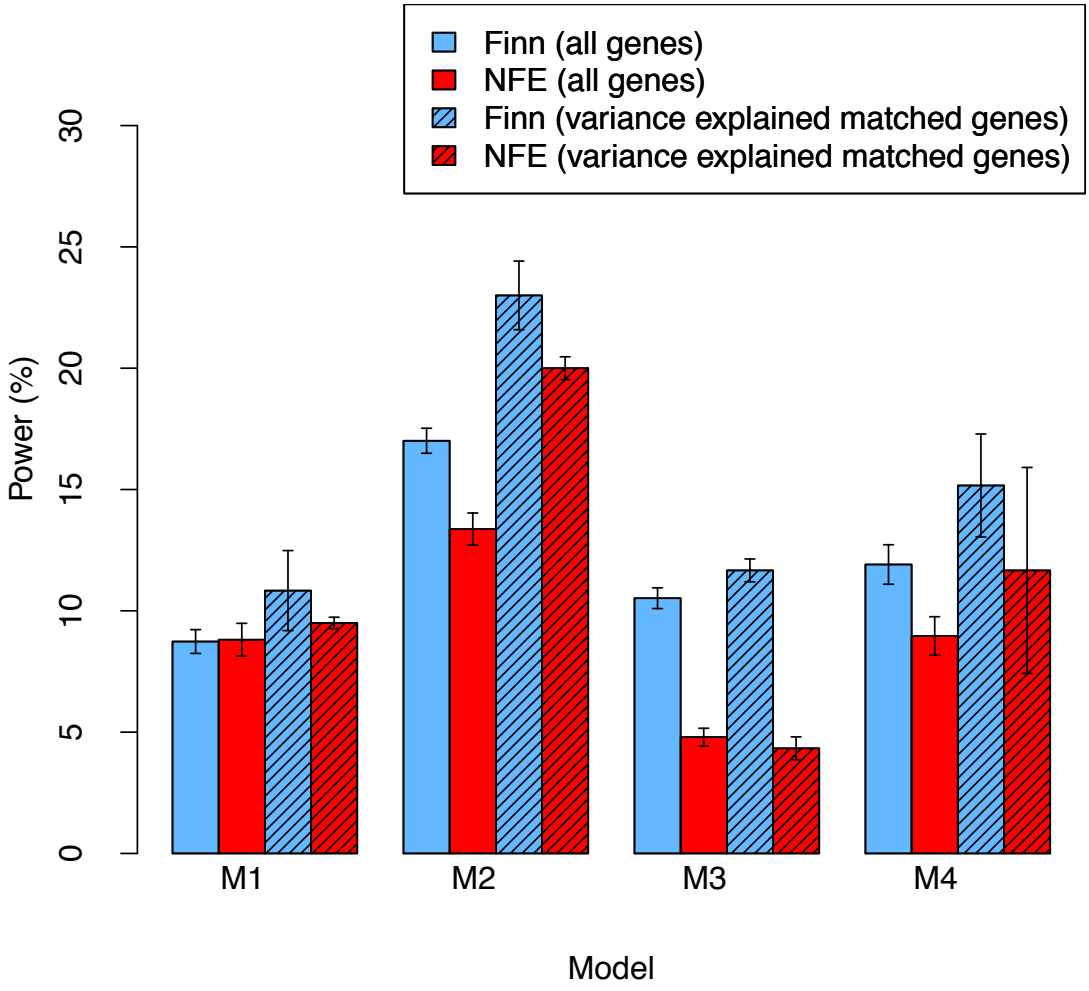

Figure S4.14 Power of SKAT-O test in 30,000 Finns or 30,000 NFEs under four different disease models, either for all genes, or for a set of genes sampled by matching the variance explained in the Finns and the NFEs.

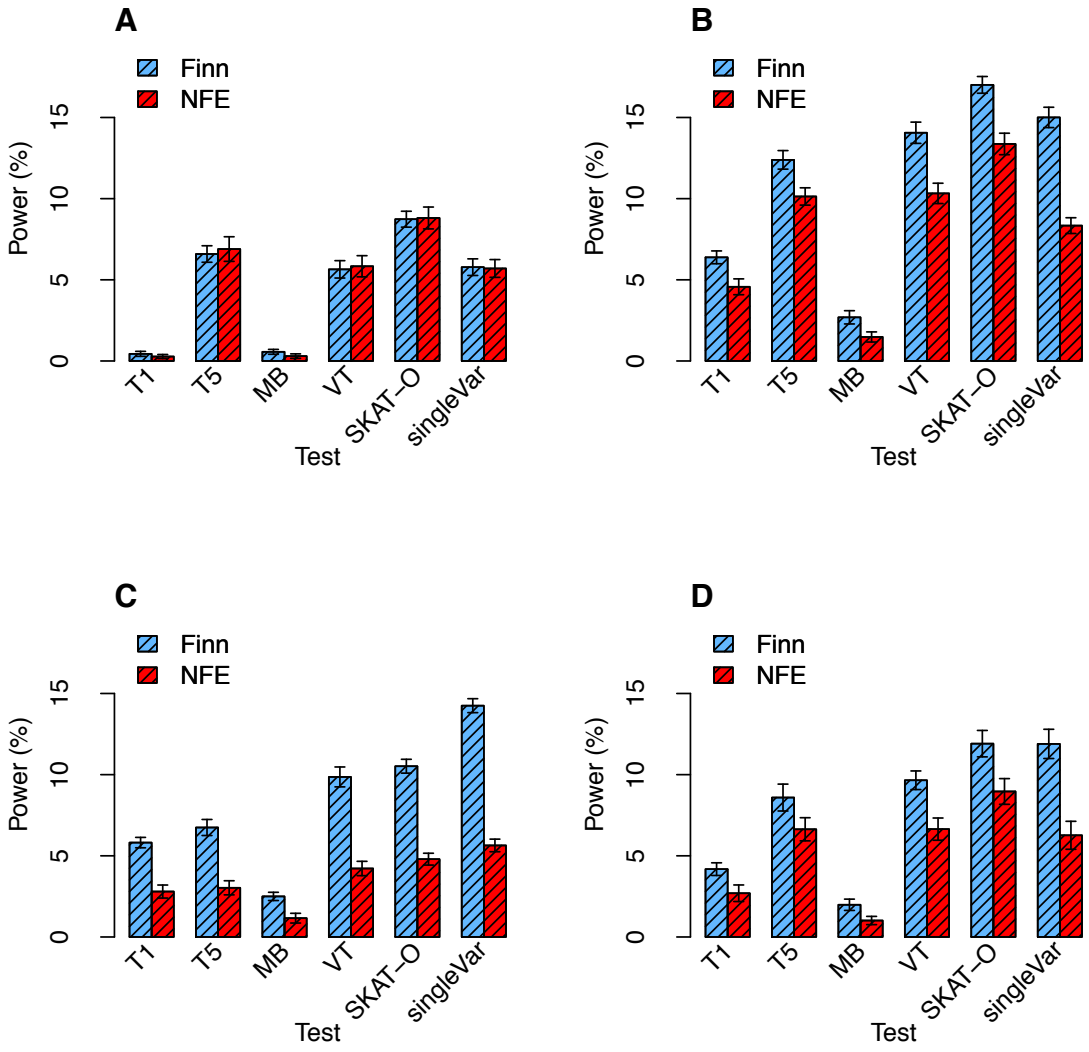

Figure S4.15: Power of exome sequencing studies in 30,000 Finns vs. 30,000 NFEs. (A) M1 (τ=0); (B) M2 (τ=0.5); (C) M3 (τ=1); (D) M4 (τ randomly sampled from 0, 0.5, and 1 for each effect gene). We simulated a quantitative trait (h2 = 80%) for which aggregated coding variation in 1,000 genes explains the total heritability. Models M1-4 were generated by varying the degree of coupling (τ) between a causal variant's phenotypic effect and the strength of purifying selection against that variant. We implemented five gene-based tests (T1, T5, MB, VT, SKAT-O) in addition to the single variant tests (singleVar) (see Methods).

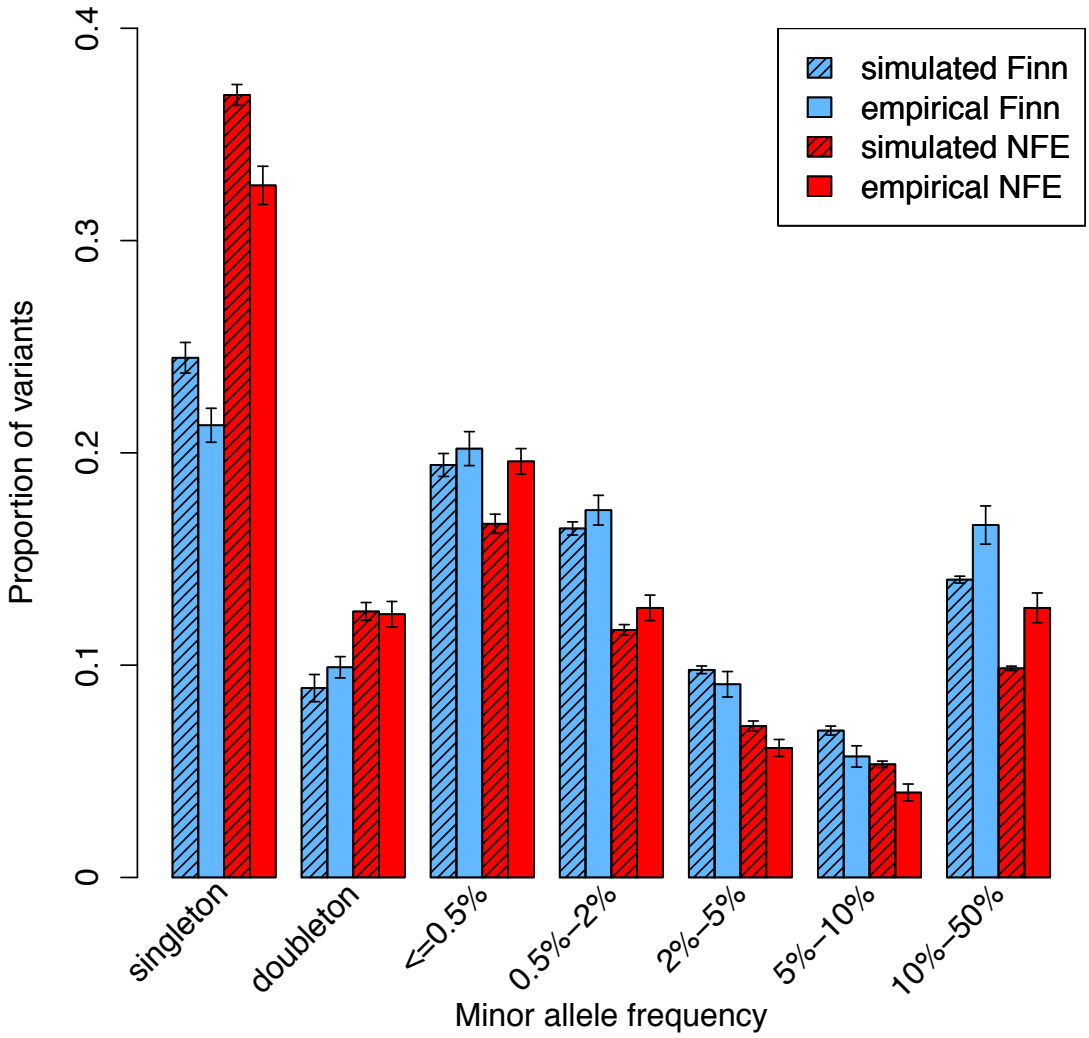

Figure S4.16: Agreement of empirical allele frequency spectra with the modeled spectra of exome chip data.

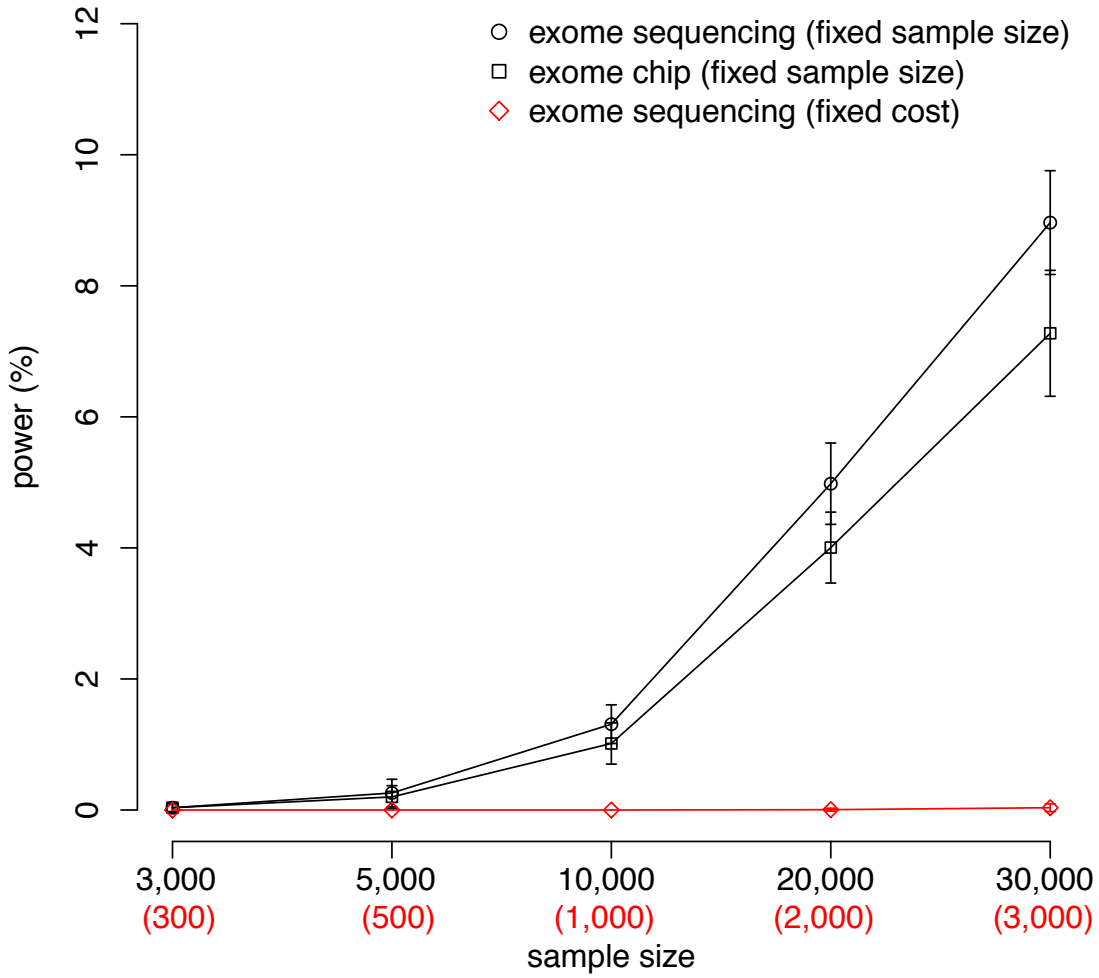

Figure S4.17: Power of exome chip study *vs* exome sequencing study in the NFEs under M4 using SKAT-O test. As different genes are likely to have different pleiotropic effects and are therefore exposed to different strengths of purifying selection, M4 is generated to represent a potentially more realistic scenario where τ (the degree of coupling between a causal variant's phenotypic effect and the strength of purifying selection against that variant) is randomly chosen among 0, 0.5 and 1 for each effect gene. The top two lines show power comparison at a fixed sample size; the bottom two lines show power comparison at a fixed cost (and thus only a tenth of the samples were sequenced).

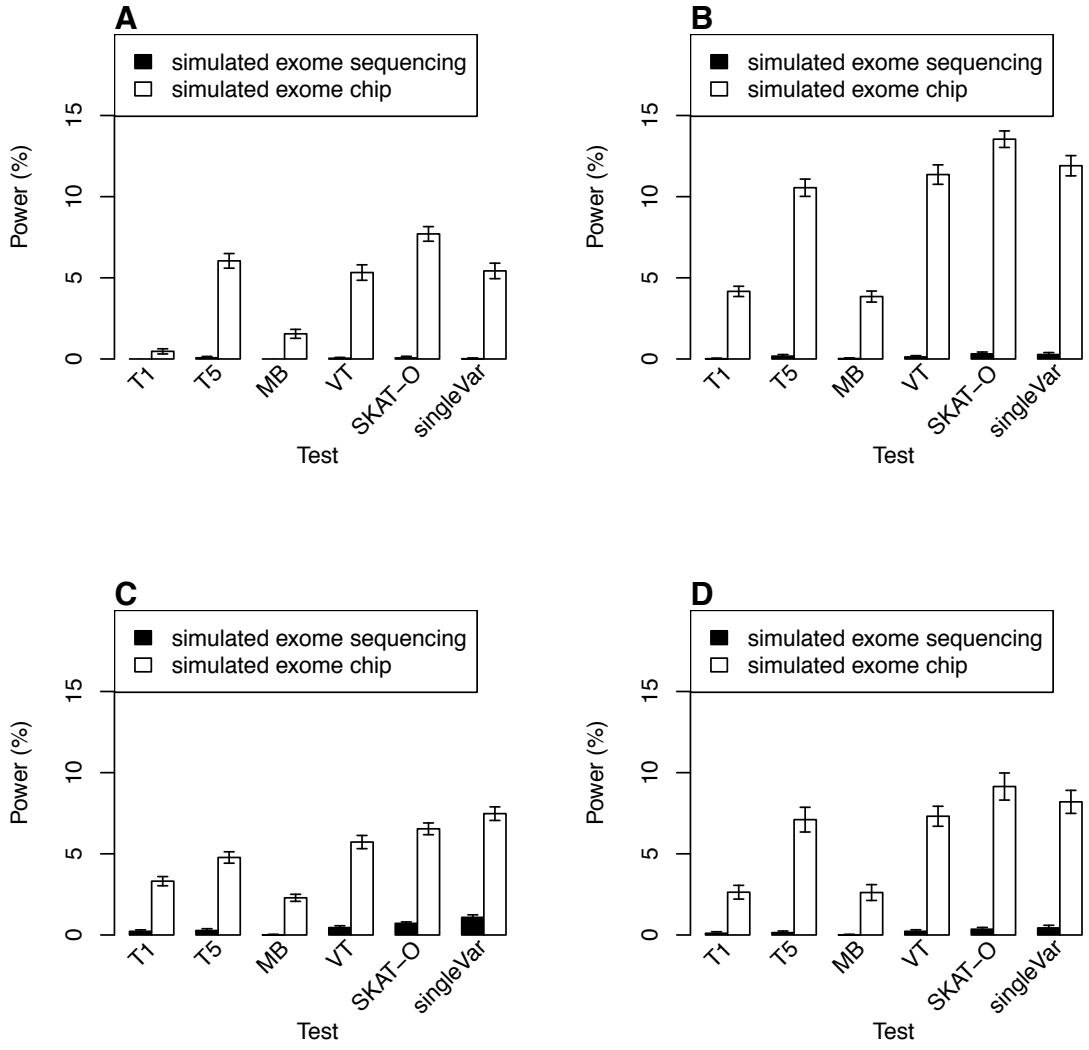

Figure S4.18: Power of exome chip study (N=30,000) *vs* exome sequencing study (N=3,000) in the Finns under four different disease models. (A) M1: τ=0; (B) M2: τ=0.5; (C) M3: τ=1; (D) M4: τ randomly sampled from 0, 0.5, and 1 for each effect gene.

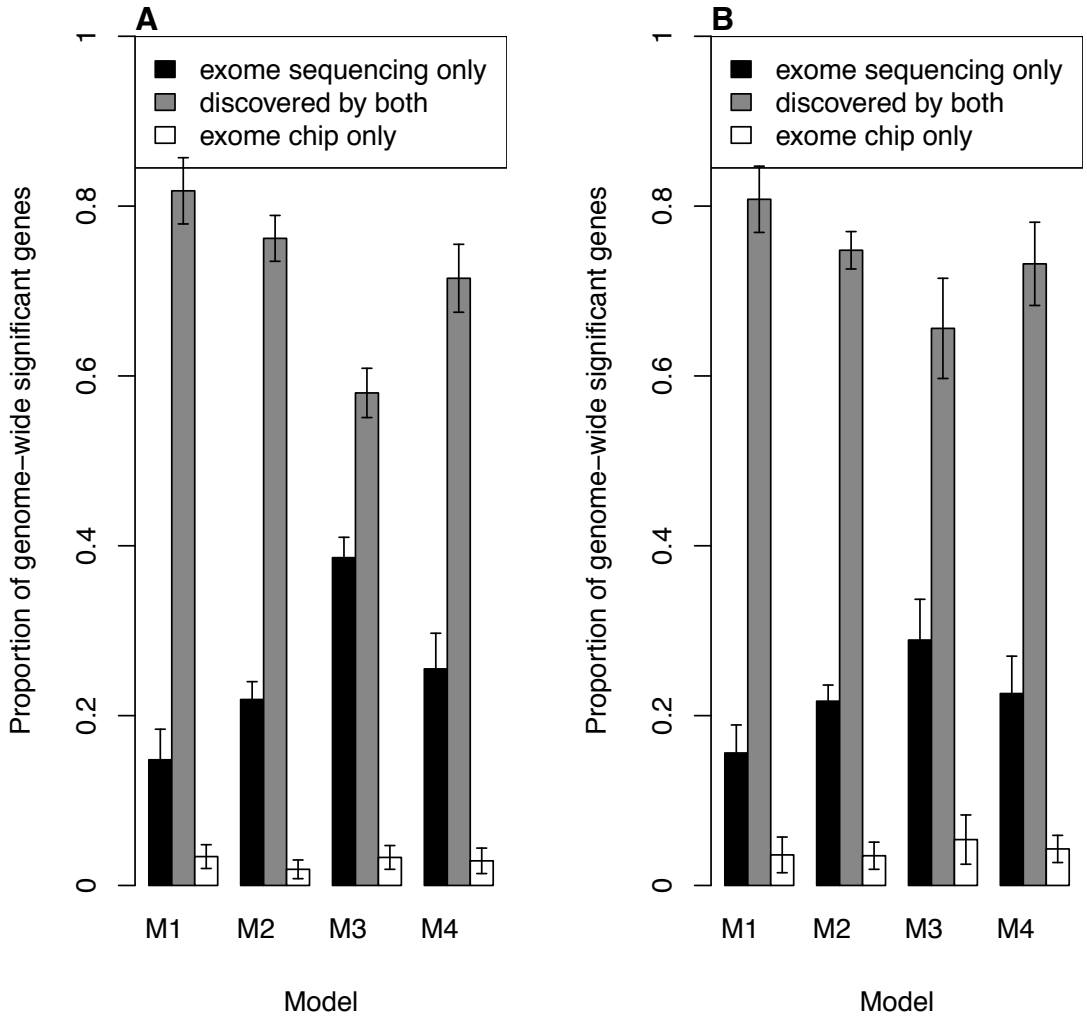

Figure S4.19: Proportion of genes detected by exome sequencing (N=30,000) only, or by exome chip (N=30,000) only, or by both (using SKAT-O test) under four different disease models (M1: τ=0; M2: τ=0.5; M3: τ=1; M4: τ randomly sampled from 0, 0.5, and 1 for each effect gene). As τ gets larger, the proportion of genes detected by exome sequencing only increases. (A) Results in Finns; (B) results in NFEs.

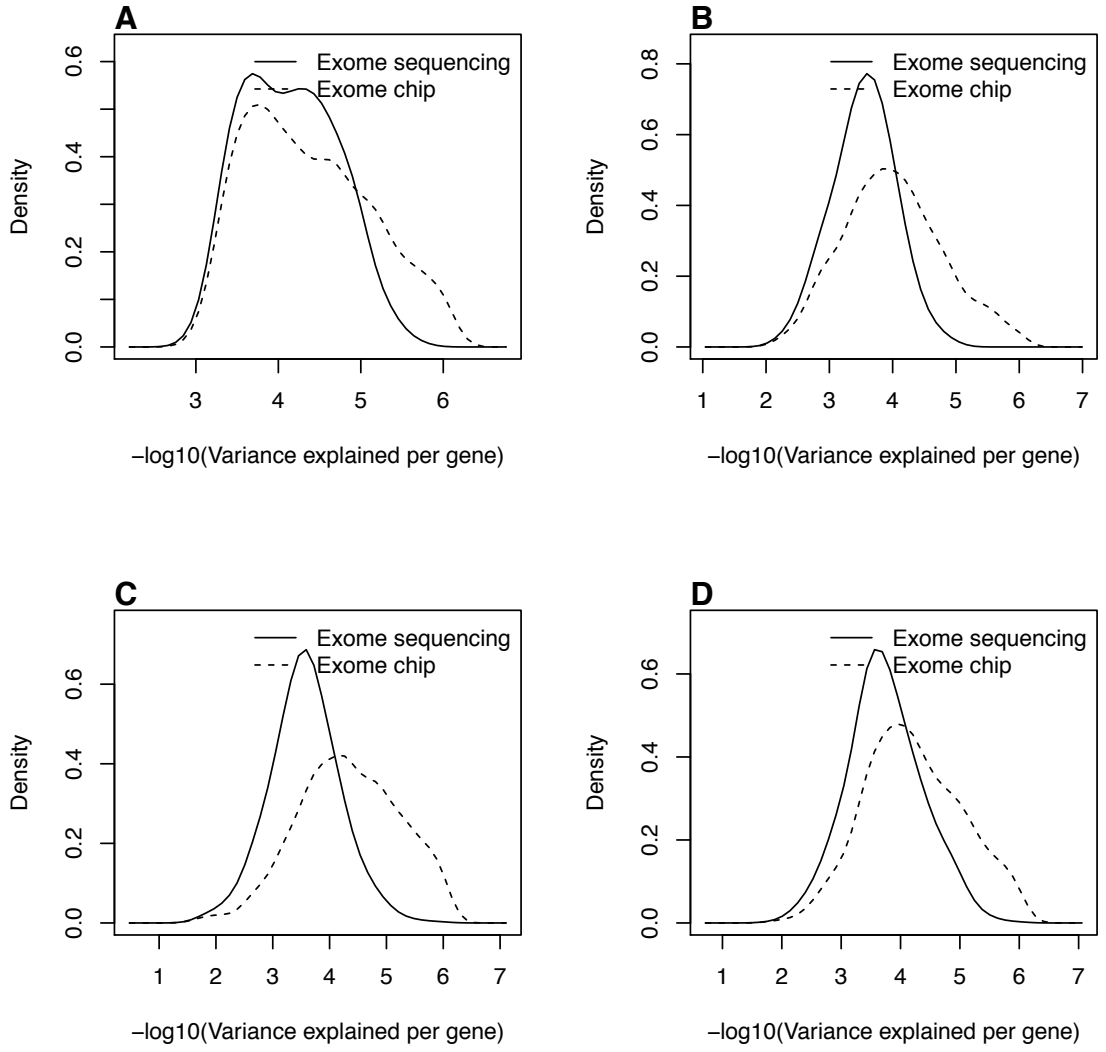

Figure S4.20: Distribution of variance explained per gene by variants with MAF below 5% in exome sequencing (solid line) or exome chip (dashed line) data of 30,000 Finns, under four different disease models. (A) M1 (τ=0); (B) M2 (τ=0.5); (C) M3 (τ=1); (D) M4 (τ randomly sampled from 0, 0.5, and 1 for each effect gene).

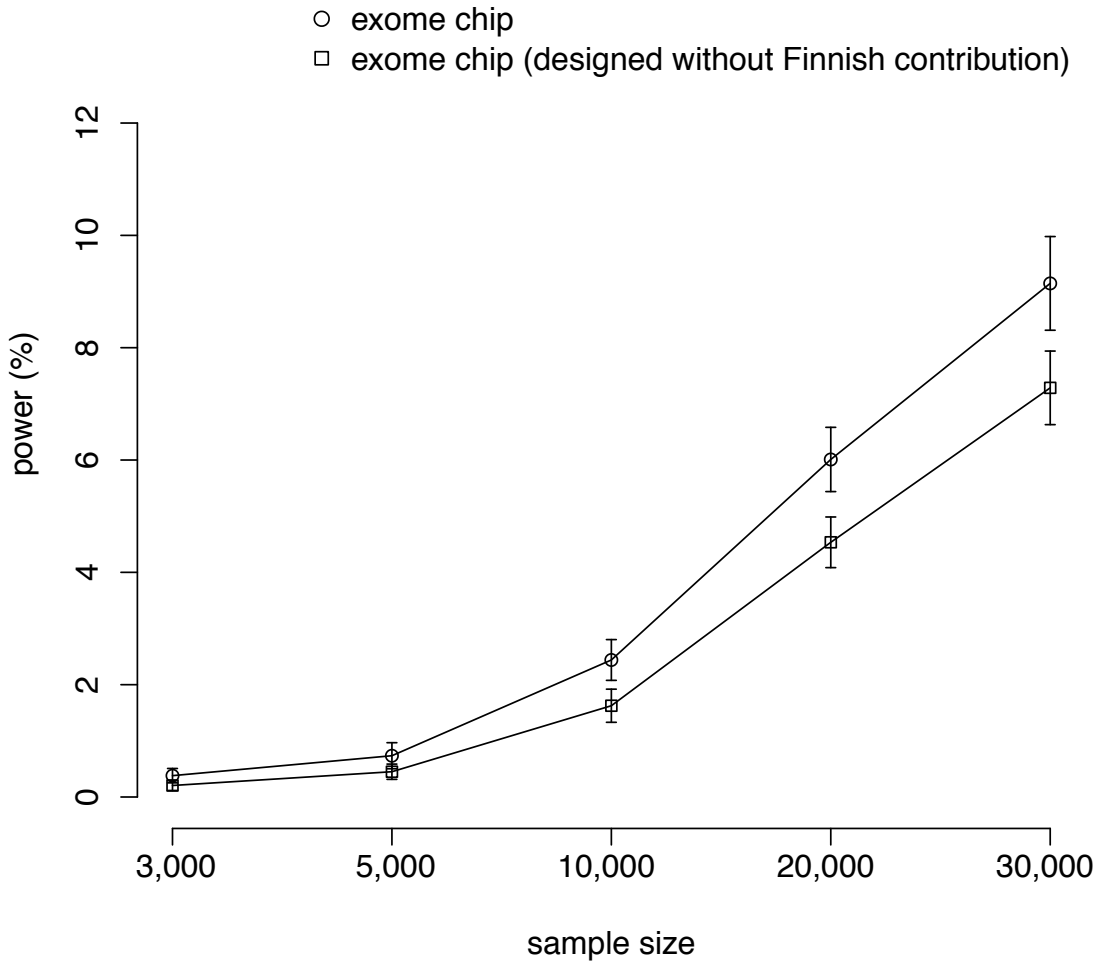

Figure S4.21: Power of two different exome chips in the Finns under M4 (τ randomly sampled from 0, 0.5, and 1 for each effect gene) using SKAT-O test. One chip design resembles that of the actual exome chip design (top line); the other chip design uses NFE samples only with no contribution from Finnish samples (bottom line).

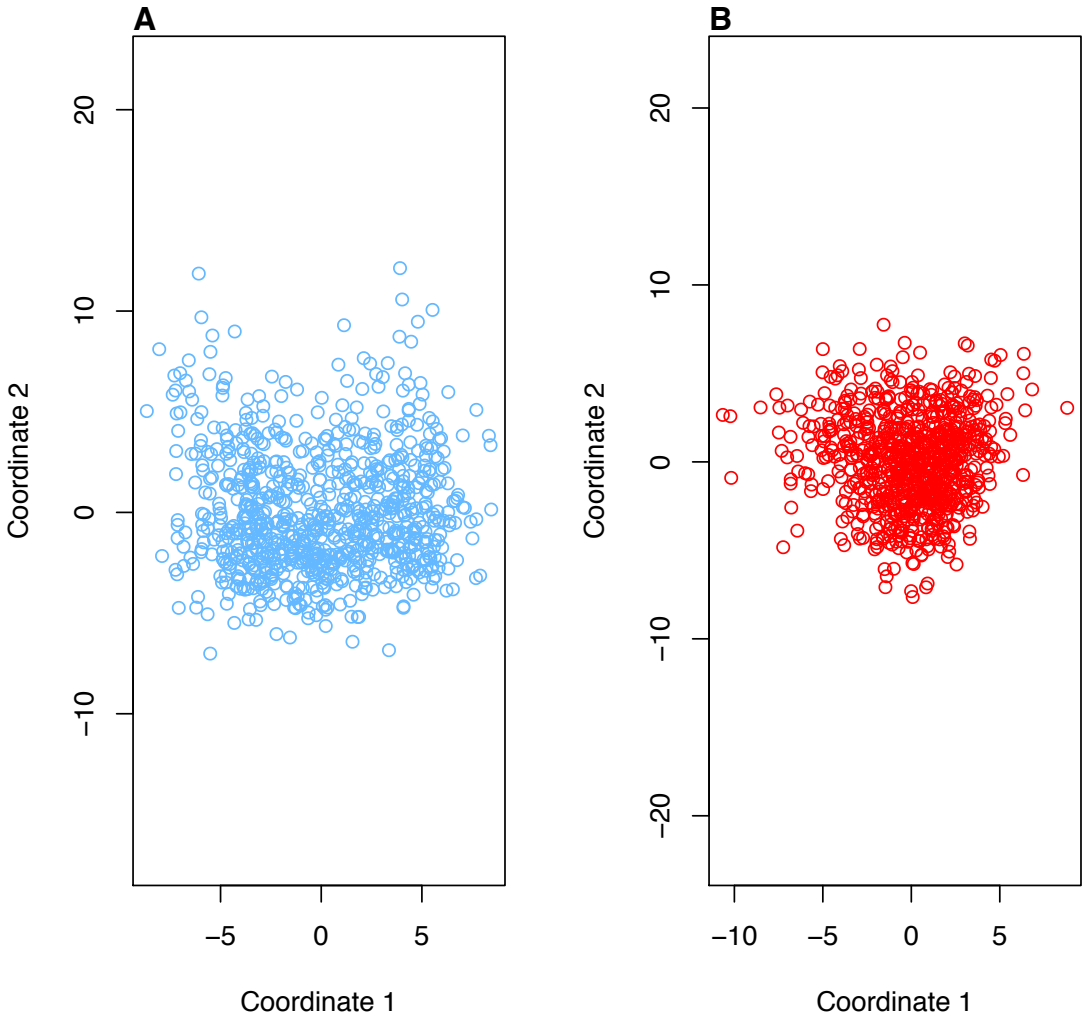

Figure S4.22: MDS plots of whole exome sequenced samples from GoT2D project. (A) Finns (N=843); (B) NFEs (N=820).

Table S4.1: Variants found in both samples tend to have higher allele counts in Finns

Variants	Finn-NFE		Finn-Finn		NFE-NFE	
	Difference	p value	Difference	p value	Difference	p value
Synonymous	0.415	0.000733	0.00451	0.911	0.0059	0.845
Missense	0.505	1.43e-06	0.00868	0.777	0.000157	0.994

For Variants shared between 250 Finns and 250 NFEs, their allele counts tend to be higher in Finns (paired t-test). As controls, we also checked allele counts for variants shared between 250 Finns and another 250 Finns, as well as between 250 NFEs and another 250 NFEs.

Table S4.2: Birth place distribution of FUSION samples

Birth place	No. of samples
UUSIMAA, UUDENMAAN / NYLAND	27
TURKU-PORI, TURUN JA PORIN / ABO-BJORNEBORG, ABO-OCH-BJORNEBORG	92
HAME, HAMEEN / TAVASTEHUS	105
KYMI, KYMEN / KYMMMENE, VIBORG, VIIPURI	49
MIKKELI, MIKKELIN / SAINT MICHEL	61
POHJOIS-KARJALA, POHJOIS-KARJALAN / NORRA-KARALEN, NORRA KARENS	52
KUOPIO, KUOPION / KUOPIO	148
KESKI-SUOMI, KESKI-SUOMEN / MELLERSTA-FINLAND	76
VAASA, VASAAN / VASA, WASA	125
OULU, OULUN / ULEABORG	41
LAPPI, LAPIN / LAPPLAND	13
KARJALA, VIIPURI (area formerly part of Finland)	54
Total	**843**

Table S4.3: Three different models of Finnish population history

Parameters	Class 1 Model[a]	Class 2 Model[b]	Class 3 Model[c]
Bottleneck size	200-4000	1000	1000
Bottleneck time	1.5-3.5ky ago	2.5ky ago	2.5ky ago
Growth rate (per generation)	2.5-10%	5-10%	Slow phase:0.5-5% Fast phase: 8-30%
Gene flow into Finns	0	1-5%	0.5-7%
Minimal -log(P(data|model))	1419	426	267

[a]Founding bottleneck event followed by exponential growth of constant growth rate, with no gene flow between NFEs and Finns

[b]Founding bottleneck event followed by exponential growth of constant growth rate, with gene flow from NFEs into Finns

[c]Founding bottleneck event followed by a slow growth phase and a fast growth phase, with gene flow from NFEs into Finns